Business Intelligence en el ámbito Académico

Proyecto Fin de Carrera
Ingeniería Técnica en Informática de Gestión
Universidad Carlos III de Madrid

Catalina Nájera Bellón

Tutor: Anabel Fraga Vázquez

Leganés, 26 de Octubre de 2015

A mi familia, por vuestro apoyo incondicional,

A Javi, simplemente, por hacerlo posible.

Sabeis lo que significaba para mi presentar este proyecto.

Gracias.

1 TABLA DE CONTENIDO

Business Intelligence en el ámbito Académico

2 PRESENTACIÓN

Cada vez manejamos mayor cantidad de información que poco a poco se va haciendo imposible de abordar. Ya no solo las grandes empresas, sino las PYMES y pequeñas empresas van generando la necesidad de poder controlar y organizar dicha información, para poder posicionarse en el mercado.

Para las grandes compañías esto no supone ningún problema, porque dispone de recursos y herramientas de Bussiness Inteligence y Big Data que les ayudan en esa labor. Pero, ¿y las PYMES? ¿existe alguna herramienta de BI en el mercado que no suponga de una gran inversión?

Por otro lado, existen infinidad de recursos en Internet que podrían sernos de gran ayuda para tomar decisiones estratégicas en nuestra empresa si se organizan y explotan de la forma adecuada.

Este proyecto surgió para dar respuesta a la primera pregunta, y se desarrolló apoyándose en los recursos gratuitos ofrecidos por el Ministerio de Educación.

3 OBJETIVOS

El presente proyecto se concibió con un objetivo inicial que ha ido evolucionando, adaptándose a los recursos e información disponibles:

El objetivo principal es estudiar la viabilidad de desarrollar proyectos de Data Warehouse en PYMES, teniendo en cuenta el presupuesto limitado de las mismas. Para ello, se deberá estudiar las diferentes herramientas de código libre que hay en el mercado y seleccionar la más apropiada, así como evaluar los costes de desarrollo y mantenimiento.

Para cumplir con dicho objetivo, y ya que no se dispone de una fuente de datos de una PYME para desarrollarlo, surge el segundo objetivo: aprovechar los recursos que ofrece el Ministerio de Educación para conocer las posibilidades de la herramienta. Con esta fuente de datos, se deberá transformar información en **conocimiento**.

Además, muchos de estos recursos gratuitos subidos a Internet, no sólo por el Ministerio de Educación, sino por innumerables entidades, como el Instituto Nacional de Estadística, están generados con un programa estadístico especializado, llamado PC-AXIS, que supone un estándar europeo en cuanto a la difusión de datos. El formato de estos archivos es similar a la de un cubo dimensional, con una serie de metadatos al inicio. Se deberá desarrollar un script genérico capaz de transformar estos archivos en ficheros de texto con un formato apropiado para poder cargar en una tabla de un modelo relacional.

4 ESTADO DEL ARTE

Como ya se ha comentado anteriormente, uno de los requisitos del proyecto es que la implantación del DWH sea al mínimo coste posible, por lo que el primer punto es seleccionar una herramienta Open Source.

Como apoyo en esta tarea, nos podemos fijar en el Cuadrante Mágico de Gartner[1][01] para plataformas de BI y analíticas que se publicó en febrero de 2015:

Ilustración 1: Cuadrante Mágico de Gartner para plataformas Analíticas y BI

[1] Análisis anual presentado por la empresa consultora y de investigación de las tecnologías de la información Gartner Inc.

Se puede observar que los líderes indiscutibles siguien siendo, año tras año las plataformas comerciales habituales:

- SAP Business Object [2]
  La combinación de SAP Business Objects y SAP NetWeaver BW cuenta con la mayor parte del mercado de plataformas de BI.
  La principal característica de la estrategia BI de SAP es su capacidad de integración entre las diferentes aplicaciones empresariales, más aun si se tiene en cuenta que la compañía cuenta con la mayor base instalada entre las plataformas de BI. Además el liderazgo en el sector ERP le permite aumentar sus ventas mediante estrategias de venta cruzada.

- IBM Cognos [3]
  Cognos 10 es la solución BI de IBM. La compañía combina un completo software, hardware y servicios en un mercado con una oferta coordinada, aportando una visión de BI unificada, tanto en el área de análisis y gestión del rendimiento.
  Las razones principales por las que los clientes seleccionan IBM son la funcionalidad, la facilidad de uso para los usuarios finales, y el acceso e integración de datos.

- Oracle BI [4]
  Su principal componente es Oracle Business Intelligence Enterprise Edition (OBIEE). Durante 2012, Oracle anunció la adquisición de Endeca con el objetivo de extender sus posibilidades en el área de capacidades analíticas y Data Discovery.
  Oracle ha adoptado tradicionalmente una estrategia de adquisición. Por este motivo, ha de entenderse que los principales esfuerzos de la compañía tienen como objetivo alinear sus productos en una única plataforma, con objeto de proveer a sus clientes de una oferta unificada.

- MicroStrategy [5]
  Los clientes de MicroStrategy mencionan que las principales razones de su selección son la funcionalidad, rendimiento y la capacidad de gestionar grandes volúmenes de datos. Permitiendo implantaciones para gran número de usuarios y complejos requerimientos de negocio.
  MicroStrategy tiene una visión centrada en dotar al usuario de funcionalidades de alto valor, en particular para la movilidad, siendo uno de los primeros proveedores en invertir en implementación de aplicaciones de BI en dispositivos móviles. Inicialmente en Research In Motion (RIM) de los dispositivos BlackBerry, para continuar apostando en estos momentos por la plataforma de Apple, con desarrollos para iPhone y iPad, como para dispositivos Android.

- Microsoft [6]
  Microsoft ofrece un conjunto competitivo de capacidades de BI y una atractiva política de precios y beneficios que atraen a los desarrolladores de Microsoft y su canal de distribuidores independientes.
  La compañía ha invertido en la construcción y mejora de las capacidades de BI en tres de las ofertas de sus productos principales: Microsoft Office (especialmente Excel), Microsoft SQL Server y Microsoft SharePoint. De esta forma Microsoft prácticamente garantiza que se siga

adoptando su oferta de BI, especialmente en organizaciones con una infraestructura de información en Microsoft.

En el Cuadrante Mágico de Gartner también cabe destacar la posición de liderazgo de Tableau[9] y Qlik[10], gracias a su agilidad y a su bajo coste, aunque no son herramientas que ofrezcan una solución BI completa propiamente dicha.

Respecto a las principales herramientas Open Source, Pentaho[7] aparece en el cuadrante de "nicho", ya que está dedicado a un mercado más pequeño, y Jaspersoft de Tibco en el de "visionarios", ya que tienen una visión de hacia donde se dirige el mercado, aunque no la esté ejecutando correctamente en el momento.

4.1 PENTAHO VS JASPERSOFT

Pentaho[7] es una completa solución Business Intelligence desarrollada bajo la filosofía Open Source para la gestión y toma de decisiones empresariales. Es una plataforma compuesta de diferentes programas que ofrecen soluciones para la gestión y análisis de la información, que incluyen ETL, entorno web, análisis multidimensional OLAP, presentación de informes, minería de datos y creación de cuadros de mando para el usuario.

Cuenta con una comunidad basada en desarrollo, que realiza mejoras constantes y extensiones en la plataforma.

Hitachi Data Systems (HDS) anunció, en junio de 2015 la adquisición del software de Pentaho. HDS tiene una historia de compromiso con el software de código abierto, por lo que se espera que dicha adquisición no afecte negativamente a la versión Community de Pentaho.

Por otro lado, al igual que Pentaho, Jaspersoft[8] compone un conjunto de herramientas Open Source que permite a las organizaciones generar información basada en sus datos de administración para la evaluación y toma de decisiones. En este caso fue TIBCO quien la adquirió en mayo de 2014.

Jasperfsoft también dispone de una comunidad de código abierto, cuyos miembros demuestran una gran actividad.

Son muchos puntos en común lo que tienen estas dos plataformas, y en internet existen infinidad de comparaciones cuya balanza torna de un lado a otro según el autor de la misma.

	pentaho	JASPERSOFT	
Informes		x	Los informes son el punto fuerte de Jaspersoft, con un diseño "pixel-perfect" interactivo.
Cuadros de Mando	x		En Jaspersoft sólo está disponible en la versión Enterprise. Además, Pentaho la supera en diseño y funcionalidad.
Plugins	x		
Compatibilidad móvil		x	Pentaho sólo en edición empresarial.
Documentación		x	
ETL	x		Jaspersoft utiliza Talend, sin incluir algunas funcionalidades.
Análisis OLAP	x	x	Ambas utilizan Mondrian.
Comunidad			La comunidad de Jaspersoft es más activa.
Minería de datos	x	x	
Big Data	x	x	
Integración	x		

Tabla 1: Pentaho vs Jaspersoft

Ambas plataformas cuentan también con grandes clientes, Pentaho por ejemplo con Telefónica en un programa de acción social, o Mozilla, y Jaspersoft con la Universidad del Este de Londres, o Ericsson, entre otros muchos.

En general, se podría decir que el punto fuerte de Pentaho es la ETL y la integración de datos, y el de Jaspersoft el diseño de informes.

Frente a esta difícil decisión, se ha dado prioridad a la robustez e integración de la plataforma, una mejor herramienta ETL y mayor experiencia por parte de desarrollo, frente a una mejor presentación de la información, por lo que se ha optado por Pentaho.

4.2 TECNOLOGÍAS UTILIZADAS

4.2.1 Vmware Workstation 10

VMware Workstation[11] es una de las herramientas de virtualización más potentes en la actualidad, con la que se puede recrear las características técnicas de un ordenador e instalar un sistema operativo sin necesidad de realizar particiones. Estas instalaciones pueden trasladarse a cualquier otro equipo que también posea Vmware Workstation.

Soporta Ubuntu, sistema con el que se pretende desarrollar el proyecto.

4.2.2 Ubuntu Server 12.04

Ubuntu[12] es un sistema operativo basado en GNU/Linux que se distribuye como software libre, el cual incluye su propio entorno de escritorio denominado Unity.

Está orientado al usuario novel y promedio, con un fuerte enfoque en la facilidad de uso y en mejorar la experiencia de usuario.

4.2.3 CPAN para Perl

Herramienta que gestiona los paquetes Perl desde el repositorio CPAN[13].

Como módulo adicional, se incluye **PC::Axis**[14] para el tratamiento de ficheros estadísticos.

4.2.4 MySQL

MySQL[15] es un gestor de bases de datos relacionales, multihilo y multiusuario, que se ofrece bajo la licencia GNU GPL para uso en productos no privativos.

4.2.5 Emma

Emma[16] es una sencilla herramienta para el manejo de BBDD MySQL. Permite la exportación de resultados en formato CSV y la edición de varias bases de datos de forma simultánea.
Resulta bastante intuitivo y evita el tratamiento con la línea de comandos.

4.2.6 Oracle Java 7**[17]**

Requisito para la instalación de las herramientas de Pentaho.

4.3 PENTAHO[7]

Conjunto de herramientas integradas para generar inteligencia empresarial, a partir de informes, minería de datos, ETL, etc.

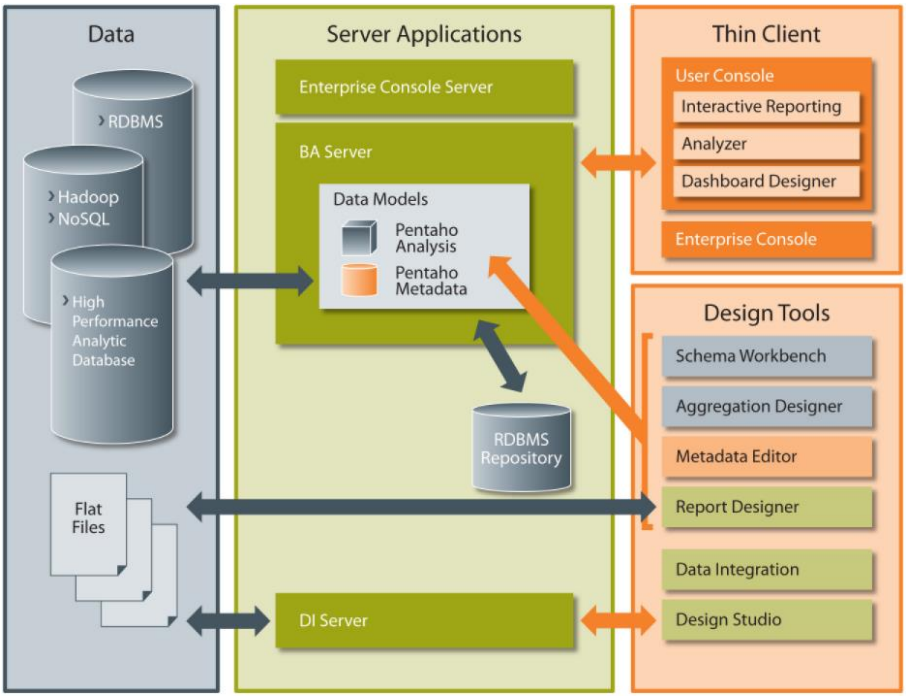


Ilustración 2: Componentes Pentaho

La suite de Pentaho ofrece dos versiones, la edición Enterprice, y la edición Community.

La versión Enterprice tiene características y servicios de soporte que no se encuentran en la Community y se obtiene a través de suscripciones anuales.

En la siguiente tabla se resumen los productos y plug-in más populares de Pentaho:

Producto	Edición	Tipo	Descripción
Pentaho BI Platform[20]	EE, CE	Aplicación Servidor	Comúnmente conocido como Plataforma BI, y recientemente rebautizada como Plataforma Business Analytics , conforma la pieza de software principal que aloja contenidos creados tanto en el propio servidor a través de plug-in como archivos publicados en el servidor a partir de aplicaciones cliente. Incluye funciones para la gestión de la seguridad, la ejecución de informes, presentación de cuadros de mando, reglas de negocio, análisis OLAP etc. La Plataforma BI corre en Apache Java Application Server.
Pentaho Analysis Services (Mondrian)[36]	EE, CE	Aplicación Servidor	Pentaho Analysis Services, también llamado Mondrian, es un servidor OLAP de código abierto (procesamiento analítico en línea), escrito en Java. Es compatible con MDX (expresiones multidimensionales) lenguaje de consulta y XML para las especificaciones de interfaz de Anasilys y olap4j. Puede leer de SQL y otros orígenes de datos e información agregada en memoria caché. Mondrian se puede ejecutar por separado, pero siempre viene integrado con la plataforma BI de Pentaho, tanto en la versión EE como en la CE.
Pentaho Dashboard Designer (PDD)	EE	plug-in Servidor	Plug-in comercial para los suscriptores de la edición Enterprice (EE). Permite la creación de cuadros de mando, cuyo objetivo es el de proporcionar una visión centralizada de los indicadores clave de rendimiento (KPI) s y otros movimientos de datos de negocio, permitiendo a los usuarios monitorizar y tomar decisiones.
Pentaho Data Integration (PDI)[18]	EE, CE	Aplicación Escritorio	Pentaho Data Integration, comúnmente llamado Kettle, es una herramienta ETL que permite a los usuarios la construcción de trabajos y transformaciones.
Pentaho Big Data[37]	EE, CE	PDI plug-in	Pentaho Big Data es una herramienta de integración de datos basada en Pentaho Data Integration. Permite la ejecución de trabajos de ETL dentro y fuera de los entornos Big Data, como Apache Hadoop o distribuciones Hadoop como Amazon, Cloudera, EMC Greenplum, MapR y Hortonworks . También es compatible con fuentes de datos no SQL como MongoDB y HBase.

Pentaho Report Designer[19]	EE, CE	Aplicación Escritorio	Pentaho Report Designer es un editor de informes basado en bandas. Entre sus principales características se incluyen el uso de subinformes, tablas y gráficos. Se puede consultar y utilizar los datos de muchas fuentes, incluyendo SQL, MDX, acceso a datos de la Comunidad, scripting, definiciones de tablas estáticas, etc.
Pentaho Data Mining (Weka)[38]	EE, CE	Aplicación Escritorio	Pentaho Data Mining utiliza Weka para buscar patrones de datos. Weka contiene algoritmos de aprendizaje automático para un amplio conjunto de tareas de minería de datos. Contiene funciones para el procesamiento de datos, análisis de regresión, métodos de clasificación, análisis de conglomerados y visualización. Sobre los patrones descubiertos, los usuarios pueden predecir las tendencias futuras.
Pentaho Metadata Editor (PME)[39]	EE, CE	Aplicación Escritorio	Pentaho Metadata Editor se utiliza para crear modelos de negocio y actúa como una capa abstracta con las fuentes de datos subyacentes. Los modelos de metadatos resultantes son utilizados por Pentaho Reporting, Saiku y otros plug-in de informes de Pentaho para crear informes en el servidor de BA sin necesidad de utilizar cualquiera de las otras aplicaciones de escritorio externos.
Pentaho Aggregate Designer (PAD)	EE, CE	Aplicación Escritorio	Aggregate Designer opera en Pentaho Analysis (Mondrian) para generar precálculos y respuestas agregadas para agilizar el trabajo de análisis y consultas MDX ejecutadas contra Mondrian. Después de usar el software para generar estas tablas agregadas, el esquema XML original Mondrian que describe el cubo OLAP se modifica para hacer referencia a los resultados precalculados.
Pentaho Schema Workbench (PSW)	EE, CE	Aplicación Escritorio	Pentaho Schema Workbench proporciona una interfaz gráfica en el diseño de cubos OLAP para Pentaho Analysis (Mondrian). El esquema creado se almacena como un archivo XML.

Pentaho Design Studio (PDS)[**40]**	EE, CE	Aplicación Escritorio	El servidor de Pentaho BA soporta scripts XML especiales llamados xactions, para implementar la lógica de negocio y otras formas de automatización en la plataforma. Design Studio es una versión modificada del entorno de desarrollo de Eclipse con un plug-in diseñado para comprender los componentes soportados por los scripts xaction. Estos scripts son muy poderosos, y útiles, pero a veces resultan difíciles de implementar. Por eso los desarrolladores están empezando a utilizar transformaciones de Pentaho Data Integration para llevar a cabo este tipo de tareas. Las transformaciones se pueden ejecutar directamente por el servidor de BA y depurar visualmente en Pentaho Data Integration (PDI) y están ganando rápidamente el favor de la comunidad sobre xactions.
Ctools[41]		plug-in servidor	Conjunto de herramientas de la Comunidad, concebidas para suplir las carencias en la creación de cuadros de mando de Pentaho.
Saiku[42]		plug-in Servidor	Excelente visor OLAP que proporciona al usuario final una magnifica herramienta para realizar análisis de forma fácil e intuitiva.

Tabla 2: Productos Pentaho

Para el desarrollo del proyecto se han utilizado los siguientes productos:

4.3.1 Data Integration 5.2 **[18]**

También conocido como Kettle/Spoon, es el componente de Pentaho responsable del proceso de Extracción, Transformación y carga (Load) de datos (ETL).

Tiene otras funciones como:

1. Migración de datos entre aplicaciones o bases de datos.
2. Exportación de los datos de las bases de datos a archivos planos.
3. Cargas masivas de datos.
4. Limpieza de datos.
5. Integración de aplicaciones.

Es una herramienta fácil de usar, gracias a su interfaz gráfica que evita la necesidad de escribir código.

Es compatible con una amplia gama de formatos de entrada y salida, como archivos de texto, hojas de cálculo, y motores de bases de datos comerciales y gratuitos.

Por otro lado, su capacidad de transformación le permite manipular los datos con muy pocas limitaciones.

4.3.2 Pentaho Reporting 5.0

Pentaho Reporting[19] es un conjunto de herramientas para la creación de informes de píxeles perfectos. Permite transformar datos en información significativa para los usuarios.

Puede exportar los informes en diversos formatos, como HTML, Excel, PDF, texto o informes impresos, o producir informes CSV y XML para alimentar otros sistemas.

4.3.3 Pentaho Metadata Editor 5.0 **[39]**

Herramienta que permite crear un modelo derivado de un DWH con definiciones orientadas al negocio, campos calculados y formateados que sirven para que los usuarios finales puedan generar informes usando herramientas como Pentaho Reporting.

4.3.4 Pentaho BI Server 5.0 **[20]**

Permite visualizar los informes, cuadros de mando o cubos OLAP a través de una interfaz web.

Gracias a Pentaho BI Server, se pueden publicar los informes desarrollados con otras herramientas en la web.

Además, proporciona una interfaz para administrar los procesos de configuración, usuarios y programación de BI.

5 METODOLOGÍA DEL PROYECTO

La construcción e implementación de un Data Warehouse es un proceso evolutivo. La experiencia en este tipo de proyectos demuestra que el éxito está en dar preferencia al seguimiento y al contacto con el usuario final, con desarrollos cortos de partes concretas, sobre un análisis exhaustivo con una solución única final.

Según el "Manifiesto ágil"[21] , definido por críticos de los modelos de mejora del desarrollo del software, dirigidos por Kent Beck, la mayor prioridad es satisfacer al cliente mediante la entrega temprana y continua de software con valor. En él, se defiende que debe aceptarse que los requisitos cambien, incluso en etapas tardías del proyecto. Los responsables de negocio y los desarrolladores deben trabajar juntos de forma cotidiana durante todo el proyecto.

Según Ralf Kimbal, considerado el inventor del Modelo Dimensional, y pionero en la Inteligencia de Negocios, el ciclo de vida de un proyecto DWH debería centrarse en el negocio y realizar entregas en incrementos significativos.

Una solución BI no debe ser algo estático y definitivo, ya que las necesidades de negocio, los usuarios y la tecnología está en constante evolución.

Es importante dar la posibilidad al usuario de analizar y obtener información de sus datos lo antes posible, para comenzar la colaboración real entre las partes, facilitando el cumplimiento de los requisitos de negocio.

Las entregas regulares ayudan a mantener el interés del cliente, involucrándolo en el desarrollo del proyecto.

La metodología utilizada sigue un proceso iterativo que sigue los siguientes pasos:

- Definición y toma de requisitos.
- Descripción Funcional.
- Análisis y Diseño Técnico.
- Construcción.
- Pruebas unitarias e integradas.
- Puesta en producción.

Ilustración 3: Fases del proyecto

5.1 DEFINICIÓN Y TOMA DE REQUISITOS

Esta fase comienza a partir de la identificación de una idea que tiene el potencial de convertirse en un proyecto dentro de la organización. Esta idea puede surgir de la necesidad de solucionar un problema, una oportunidad o amenaza del entorno, nuevas normativas, una ventaja frente a la competencia, etc.

La finalidad es establecer los objetivos de acuerdo a los requisitos del cliente.

Se puede dividir en tres tareas:

- Identificación de Requisitos. Se realiza la definición del catálogo de requisitos del DWH. Éstos pueden ser funcionales, que responden a las necesidades del negocio, o no funcionales, como los relacionados con el rendimiento, la calidad, etc.
- Estimación de Esfuerzo. Estimación en horas por roles, de las tareas que conforman la iniciativa.
- Plan de Trabajo. Calendario y recursos que se ven involucrados en el desarrollo del proyecto.

5.2 Descripción Funcional

El objetivo de esta fase es poder acordar entre el cliente y la empresa desarrolladora el alcance final del proyecto, entrando a un nivel mucho más detallado. En ella, se debe redactar un documento que cubra los siguientes aspectos:

- Objetivo que se persigue con la iniciativa.
- Historia o antigüedad de los datos a extraer.
- Frecuencia de actualización de datos.
- Software que se sustituye, si aplica.
- Origen de datos.
- Seguridad de datos. Identificación del grupo de usuarios al que está destinada la iniciativa.
- Mapa Funcional: dimensiones e indicadores que se obtienen y cómo se relacionan entre ellos.

5.3 Análisis y Diseño Técnico

En esta fase se realiza el análisis y diseño del modelo de datos, de los procesos ETL, y de los informes y cubos:

- Análisis y diseño del modelo de datos.
  Se definen las tablas de base de datos, atributos y relaciones entre ellos, que contienen las dimensiones e indicadores del proyecto. Para ello, se utiliza la herramienta Visio.
- Análisis y diseño de la ETL.
  Se definen los procesos que extraen y transforman los datos de origen, e insertan en la base de datos de destino, es decir los trabajos y transformaciones que se desarrollan en la herramienta Kettle de Pentaho.
- Análisis y diseño de informes y cubos.
  Se definen los informes de Pentaho Reporting, a nivel de formato visual y de datos, es decir, qué dimensiones e indicadores se incluyen, parámetros, filtros, cálculos, etc.

5.4 Construcción

Se fabrica, construye, e integra el DWH de acuerdo al diseño de la fase anterior:

- Construcción del modelo de datos, ejecutando los scripts de creación de BBDD.
- Construcción de la ETL, es decir, trabajos y trasformaciones de la herramienta Kettle.
- Construcción de informes, mediante la herramienta Pentaho Reporting.

5.5 PRUEBAS UNITARIAS E INTEGRADAS

En esta fase se realiza la creación y ejecución del Plan de Pruebas de todos los procesos.

Las pruebas deben ser lo suficientemente exhaustivas para tener en cuenta todas las posibles casuísticas que se presenten en la ejecución del proceso. Éstas se pueden distinguir entre pruebas de diseño y pruebas de calificación, cuyo objetivo es demostrar que el desarrollo cumple con los requisitos del cliente, plasmados en el documento funcional y técnico.

Se valida y depura el diseño, modificando el mismo si fuera necesario a la vista de los resultados.

5.6 PUESTA EN PRODUCCIÓN

Es la última fase del ciclo de vida del proyecto. En ella se prepara toda la infraestructura necesaria para que el DWH funcione en un entorno de producción, habiendo completado toda la documentación necesaria de fases anteriores.

Otro objetivo de esta fase es conseguir que el producto sea utilizado por los usuarios, dándoles el apoyo y la formación que precisen y asegurar que los beneficios alcanzados gracias al proyecto se mantengan una vez el equipo de proyecto se retire.

6 DEFINICIÓN Y TOMA DE REQUISITOS

6.1 VISIÓN Y ALCANCE

Por regla general, ninguna PYME dispone de las herramientas necesarias para facilitar la toma de decisiones a sus gerentes, apoyándose en la información que poseen, permitiéndoles supervisar, planificar y pronosticar con velocidad y precisión, la situación de sus empresas.

Esto se debe a que dichas herramientas suelen tener un alto coste que no se pueden permitir.

Otro motivo es el desconocimiento de la existencia de este tipo de herramientas en el mercado.

Hoy en día, es indispensable implantar un sistema inteligente de negocio, si se desea obtener ventaja sobre una competencia cada vez más exigente. Este sistema inteligente es lo que llamamos Business Intelligence.

Se puede denominar al Business Intelligence como el conjunto de estrategias y herramientas que permiten transformar datos en información, y la información en conocimiento.

Ilustración 4: Pirámide DIKW

El Business Intelligence permite reunir, depurar y transformar datos de los sistemas operacionales e información desestructurada, interna y externa a la compañía, en información estructurada, para su explotación directa mediante informes, cubos, etc, dando así soporte a la toma de decisiones sobre el negocio.

Con la elaboración de este proyecto se pretende construir una plataforma Business Intelligence que pueda ser aplicable a una PYME. Para ello, se recurrirá a una herramienta Open Source de la que se deberá estudiar su capacidad, potencia, facilidad de uso y limitaciones.

Como para la elaboración del proyecto no se ha podido disponer de una fuente de datos de una PYME, se incluye un segundo objetivo al proyecto: descargar diferentes estadísticas subidas a la web del Ministerio de Educación, elaboradas con el programa PC-AXIS, software estadístico estándar en la unión europea, e incorporarlas a la plataforma Business Intelligence definida en el párrafo anterior.

Además, aunque en el proyecto nos centraremos en las estadísticas sobre el número de matriculados[35], se debe desarrollar con una perspectiva genérica, es decir, deberá ser capaz de procesar cualquier estadística creada con el software PC-AXIS.

Aunque a priori parezca que estas estadísticas ya están procesadas y a punto para que el usuario pueda estudiar sus datos, analizarlos, y obtener conclusiones a partir de los mismos, hay muchas preguntas básicas a las que es imposible responder haciendo dos simples clics. De hecho, conllevaría mucho más que eso: exportar a mano las diferentes estadísticas necesarias a un formato editable, año por año. Transformar los datos, para posteriormente realizar los cálculos necesarios para realizar el análisis.

Preguntas como:

¿Cuál ha sido la universidad con mayor crecimiento en el último año?

¿Cómo ha ido evolucionando el número de matriculados en los últimos 8 años? ¿Este dato se ha visto afectado por la crisis económica?

¿Cuál ha sido la dimensión que más ha crecido en las universidades de la competencia? ¿Han focalizado su estrategia en las mujeres, en ciclos superiores?

¿Cómo ha evolucionado mi universidad frente a las que considero la competencia?

¿Cuál está siendo la rama de enseñanza tendencia en el último año?

Estoy perdiendo matriculados ¿cuáles son las dimensiones con mayor descenso?

Asimismo, como se ha mencionado anteriormente, el proceso que incorpora estas estadísticas PC-AXIS a nuestro sistema Business Intelligence, debe ser genérico, es decir, deberá transformarlas a un formato de texto separado por tabulaciones, cuya carga al sistema sea directa. Son muchas las entidades que suben sus estadísticas en este formato:

- Instituto Nacional de Estadística[24]

- Institutos Estadísticos Regionales:
  - Gobierno Vasco[25]
  - Gobierno de Canarias[26]
  - Gobierno de Aragón[27]
  - Gobierno Balear[28]

- Poder Judicial[29]

- Entidades internacionales:
  - Dinamarca[30]
  - Suecia[31]
  - China[32]

Y como las anteriores entidades, un largo etcétera.

Rara es la organización que no está interesada en incorporar datos estadísticos nacionales e internacionales, de su ámbito empresarial, dentro de su sistema de inteligencia, para poder comparar su situación frente al resto de competidores. Por ejemplo:

- Un concesionario de coches ha tenido un parón en sus ventas en los últimos meses. Podría estar interesado en comparar ese dato con las ventas nacionales para conocer si es un problema propio, o una tendencia nacional[33].

- Una compañía distribuidora de energía eléctrica y gas podría estar interesada en conocer la evolución mensual de los índices de exportación y de Importación de estos productos industriales, para saber cuál es el mejor momento para realizar las compras y ventas[34].

Partiendo de estas premisas, en el siguiente punto se definirán los requisitos funcionales y no funcionales que debe cumplir el proyecto.

6.2 IDENTIFICACIÓN DE REQUISITOS

RF-1	Preparación Máquina Virtual BI
Versión	1.0_17/03/2015
Fuente	uc3m
Dependencias	
Descripción	Implantar en una máquina virtual, un sistema Business Intelligence utilizando una herramienta Open Source, para conocer sus capacidades y limitaciones. Se deberá medir la facilidad de instalación, integración, aprendizaje en el desarrollo y distribución.
Prioridad	5
Riesgo	Alto
Precondición	
Postcondición	Máquina virtual y sistema BI Open Source a punto para ser usados.

Tabla 3: Requisito RF-1

RF-2	Elección herramienta BI
Versión	1.0_17/03/2015
Fuente	uc3m
Dependencias	RF-1
Descripción	Se deberá realizar un estudio de las herramientas BI Open Source que hay en el mercado, y elegir la más adecuada para implantar en PYMES.
Prioridad	5
Riesgo	Muy Alto
Precondición	Preparación de Máquina Virtual BI.

Postcondición	Selección de la herramienta BI Open Source más apropiada.

Tabla 4: Requisito RF-2

RF-3	Orígenes de datos múltiples
Versión	1.0_17/03/2015
Fuente	uc3m
Dependencias	RF-2
Descripción	El sistema BI deberá estar preparado para alimentarse de cualquier origen de datos, de BBDD operacionales, y fuentes internas y/o externas.
Prioridad	2
Riesgo	Alto
Precondición	
Postcondición	

Tabla 5: Requisito RF-3

RF-4	Conversión PC-AXIS en relacional
Versión	1.0_17/03/2015
Fuente	uc3m
Dependencias	RF-1
Descripción	Crear un proceso que se pueda integrar en el sistema BI, que transforme y adapte ficheros estadísticos creados con el software PC-AXIS, en ficheros planos, con campos separados por tabulación, para su posterior carga en tablas del DWH.
Prioridad	3
Riesgo	Alto

Precondición	Fichero PC-AXIS alumnos matriculados.
Postcondición	Fichero plano adaptado a tabla relacional.

Tabla 6: Requisito RF-4

RF-5	Conversión PC-AXIS genérica
Versión	1.0_17/03/2015
Fuente	uc3m
Dependencias	RF-4
Descripción	Este proceso, además de procesar las estadísticas de número de estudiantes subidas por el Ministerio de Educación, tendrá que ser capaz de admitir cualquier fichero tipo PC-AXIS.
Prioridad	1
Riesgo	Bajo
Precondición	Cualquier fichero PC-AXIS.
Postcondición	Fichero plano adaptado a tabla relacional.

Tabla 7: Requisito RF-5

RF-6	Creación de DWH número matriculados
Versión	1.0_17/03/2015
Fuente	uc3m
Dependencias	RF-1, RF-2
Descripción	Se deberá crear un DWH a partir de los datos estadísticos de número de matriculados. Es decir, un proceso ETL que extraiga, transforme y cargue los datos en el sistema, un modelo de datos, y el desarrollo de unos informes para facilitar la labor de análisis al usuario.

Prioridad	4
Riesgo	Alto
Precondición	Datos estadísticos número de matriculados.
Postcondición	Informes para el análisis.

Tabla 8: Requisito RF-6

RF-7	Informes visualmente atractivos
Versión	1.0_17/03/2015
Fuente	uc3m
Dependencias	RF-6
Descripción	Estos informes deberán apoyarse en gráficos y herramientas de visualización.
Prioridad	1
Riesgo	Bajo
Precondición	
Postcondición	Informes con objetos visuales para el análisis.

Tabla 9: Requisito RF-7

RF-8	Informes Web
Versión	1.0_17/03/2015
Fuente	uc3m
Dependencias	RF-1, RF-6

Descripción	Los informes deberán poder subirse a un portal web, integrado en el sistema BI, para que pueda ser visualizado por todos los usuarios de la compañía, o por los usuarios a los que se les permita acceder.
Prioridad	2
Riesgo	Bajo
Precondición	
Postcondición	Informes accesibles via web.

Tabla 10: Requisito RF-8

RF-9	Informes parametrizables
Versión	1.0_17/03/2015
Fuente	uc3m
Dependencias	RF-1, RF-6
Descripción	Los informes deberán disponer de parámetros de entrada con las principales dimensiones de análisis (periodo, universidad, rango de edad, ciclo formativo, rama de enseñanza, género, etc, para que el usuario pueda seleccionar el objeto de estudio a su gusto, y obtener la información deseada con muy pocos clics.
Prioridad	3
Riesgo	Medio
Precondición	
Postcondición	Informes con parámetros de entrada definidos.

Tabla 11: Requisito RF-9

RF-10	Actualización incremental
Versión	1.0_17/03/2015

Fuente	uc3m
Dependencias	RF-1, RF-6
Descripción	El sistema no debe ser estático, es decir, podrá alimentarse de los datos estadísticos que se vayan subiendo a lo largo del tiempo. Se deberá incluir un manual de operación, para facilitar esta tarea al operador responsable de la misma.
Prioridad	3
Riesgo	Medio
Precondición	
Postcondición	Sistema incremental. Manual de operación.

Tabla 12: Requisito RF-10

RF-11	Formatos de exportación
Versión	1.0_17/03/2015
Fuente	uc3m
Dependencias	RF-1, RF-6
Descripción	Los informes podrán exportarse a diferentes formatos, como HTML, PDF, CSV, etc.
Prioridad	2
Riesgo	Bajo
Precondición	
Postcondición	Informes en HTML, PDF y Excel.

Tabla 13: Requisito RF-11

6.3 ESTIMACIÓN DE ESFUERZO

En este proyecto participan 3 roles: jefe de proyecto, analista y programador, que son desempeñados por el mismo recurso, Catalina Nájera Bellón.

La labor de cada rol o perfil es:

- Jefe de Proyecto:
  El Jefe de Proyecto realiza la estimación del esfuerzo necesario para llevar a cabo el proyecto, seleccionando la estrategia de desarrollo, fijando el calendario de hitos y entregas y estableciendo la planificación del proyecto.
  Es el encargado de dirigir el proyecto, realizando las labores de seguimiento y control del mismo, revisión y evaluación de resultados y coordinación del equipo.
  Se ocupa también de la gestión y resolución de incidencias que puedan surgir durante el desarrollo así como de la actualización de la planificación inicial.
  Entre sus funciones se encuentran la elaboración de los informes de seguimiento y el archivo de la documentación de gestión del proyecto una vez que este ha finalizado.

- Analista o Analista Programador:
  La responsabilidad de los Analistas es elaborar un catálogo detallado de requisitos que permita describir con precisión el sistema de información, para lo cual mantienen entrevistas y sesiones de trabajo con los responsables de la organización y usuarios, actuando de interlocutores entre el equipo de proyecto en lo que a requerimientos se refiere.
  Estos requisitos permiten a los analistas elaborar los distintos modelos que sirven de base para el diseño, obteniendo los diseños funcionales, modelos de datos y los diseños técnicos de ETL e informes.
  Además realizan la especificación de las interfaces entre el sistema y el usuario.

- Programador
  La función del programador es construir el código que dará lugar al producto resultante, en base al diseño técnico realizado por el analista, tanto de ETL, utilizando la herramienta Kettle de Pentaho, como de informes, utilizando Pentaho Report Designer.
  Su función es también generar el desarrollo asociado a los procedimientos de migración y carga inicial de datos.
  Igualmente se encarga de la realización de las pruebas unitarias y participa en las pruebas integradas de la aplicación.

  A continuación se detalla la estimación en horas por tipo de tarea del proyecto:

6.3.1 Resumen de la estimación:

Tarea	Descripción	Total horas
Gestión Proyecto y Formación	Gestión y Seguimiento del proyecto	40
Definición y toma de requisitos	Reuniones con usuario, actas, etc	20
Diseño Funcional		20
Análisis y Diseño técnico		**68**
Análisis y diseño Modelo de datos		16
Análisis y diseño ETL		47
Análisis y diseño informes		5
Contrucción		**98**
Construcción modelo de datos	Crear tablas	3
Contrucción ETL		62
Construcción metadatos		5
Construcción de informes		27
Pruebas Unitarias e Integradas		**73**
Pruebas unitaria ETL		47
Pruebas unitaria de informes		16
Pruebas integradas ETL		5
Pruebas integradas informes		5
Implantación		**10**
Preparar paso a producción		5
Cumplimiento estandares		5
Otros		**55**
Estudio herramienta Pentaho		15
Preparación Máquina Virtual		40
	TOTAL	**384**

Tabla 14: Resumen de Estimación

Se estima que el proyecto necesitará invertir **384** horas para su desarrollo.

6.3.2 Estimación de la ETL desarrollada en Kettle

Tabla de horas para la ETL en Kettle

		Análisis y DT	Construcción	Prueba unitaria-Integrada	Total
	Complejidad	30%	40%	30%	100%
5	Muy Alta	24,0	32,0	24,0	80
4	Alta	12,0	16,0	12,0	40
3	Media	8,1	10,8	8,1	27
2	Baja	2,7	3,6	2,7	9
1	Muy Baja	1,2	1,6	1,2	4

	Complejidad	Análisis	Construcción	Prueba unitaria	Total
Script PC-AXIS to Relational	5	24	32	24	80
t_dimensiones_edad	2	3	4	3	9
t_dimensiones_rama	2	3	4	3	9
t_agregada_edad	3	8	11	8	27
t_agregada_rama	3	8	11	8	27
j_Estadisticas_Academicas	1	1	2	1	4
	TOTAL	**47**	**62**	**47**	**156**

Tabla 15: Estimación ETL

Se necesitarán 156 horas para completar la tarea del desarrollo de los procesos ETL en Kettle.

6.3.3 Estimación de los informes desarrollados en Pentaho Report Designer

Tabla de horas para los Informes en Pentaho Report Designer

		D. Tecnico	Con. FM	Construcción	Prueba unitaria-Integrada	Total
	Complejidad	10%	10%	50%	30%	100%
5	Muy Alta	8	8	40	24	80
4	Alta	4	4	20	12	40
3	Media	3	3	14	8	27
2	Baja	1	1	4	2	8
1	Muy Baja	0	0	2	1	3

	Complejidad	Análisis y DT	Con. FM	Construcción	Prueba unitaria	Total
Estadísticas Estudiantes Universitarios	3	3	3	14	8	27
Estadísticas Universidades	3	3	3	14	8	27
TOTAL Informes		**5**	**5**	**27**	**16**	**54**

Tabla 16: Estimación Informes

Se necesitarán 54 horas para la elaboración de los informes con la herramienta Pentaho Report Designer.

6.4 PLAN DE TRABAJO

A continuación se detallan el calendario y recursos que se verán involucrados en el desarrollo del proyecto, así como la estimación de costes y presupuesto. Se ha utilizado la herramienta Microsoft Project Professional 2013 para su elaboración.

6.4.1 Recursos

ESTADO DE LOS RECURSOS

Resta trabajo para todos los recursos de trabajo

Nombre	Comienzo	Fin	Trabajo restante
Jefe de Proyecto	lun 16/03/15	mié 23/09/15	0 horas
Analista	mié 25/03/15	mié 23/09/15	0 horas
Programador	lun 11/05/15	mié 16/09/15	0 horas

Tabla 17: Recursos del Proyecto

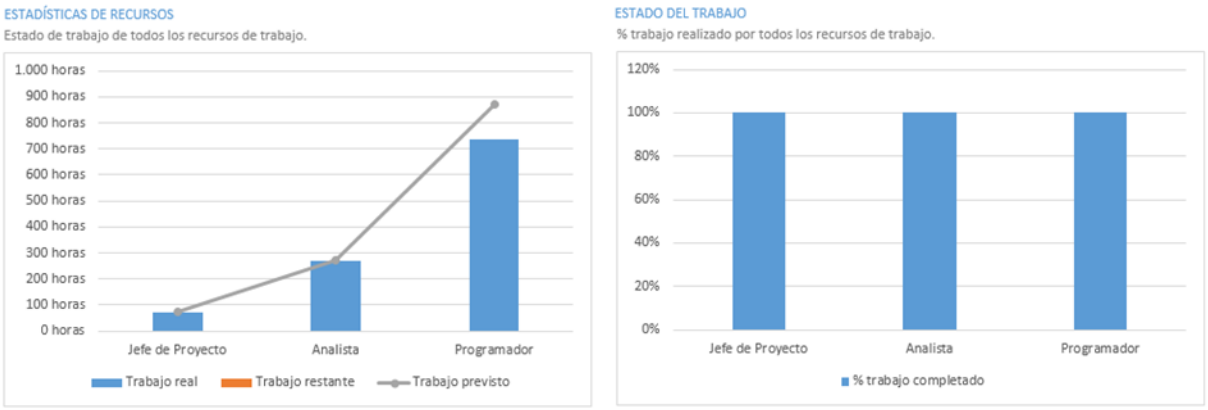


Ilustración 5: Estadísticas de Recursos y Estado del trabajo

6.4.2 Presupuesto

El presupuesto se ha dividido en los siguientes tipos de costes:

- Costes del personal involucrado en el proyecto.
- Costes de elementos software y hardware.
- Gastos de Viajes y dietas
- Material Fungible
- Gastos de documentación
- Gastos Indirectos

6.4.2.1 Costes de Personal

Nombre	Trabajo Real	Costo Real	Tasa estándar
Jefe de Proyecto	30 horas	1.050,00 €	35 €/hora
Analista	121 horas	2.420,00 €	20 €/hora
Programador	232 horas	3.248,00 €	14 €/hora
Total		**6.718,00 €**	

Tabla 18: Detalle de Costes de Personal

6.4.2.2 Coste de Software y Hardware

Producto	Coste (€)	%Uso dedicado al Proyecto	Dedicación (meses)	Periodo Depreciación	Coste Imputable
Portatil toshiba intel core 2 duo	450	100	7	60	52,5
Licencia Microsoft Oficce 2013	97	100	7	60	11,3
Licencia VMware Workstation	119	100	7	60	13,9
Licencia Ubuntu	0	100	7	60	0,0
Licencia Herramientas Pentaho, versión CE	0	100	7	60	0,0
Total					**77,70 €**

Tabla 19: Detalle de Costes Software y Hardware

6.4.2.3 Otros costes

Descripción	Coste Imputable
Gastos de Viajes y dietas	50
Material Fungible	50
Gastos de documentación	60
Gastos Indirectos	10
Total	**170,00 €**

Tabla 20: Otros costes

6.4.2.4 Presupuesto Final

Descripción	Coste (€)
Costes de Personal	**6.718,00**
Costes de Software y Hardware	**77,70**
Otros costes	**170,00**
TOTAL PROYECTO	**6.965,70 €**

Tabla 21: Presupuesto Final

El presupuesto final de este proyecto asciende a la cantidad de **6.965,70 €**.

6.4.3 Desglose de Tareas

Nombre de tarea	Duración	Comienzo	Fin	Nombres de los recursos
BI Aplicado a PYMES	**383**	**16/03/15**	**01/10/15**	
Definición y toma de Requisitos	**20**	**16/03/15**	**24/03/15**	**Jefe de Proyecto**
Visión y Alcance	5	16/03/15	17/03/15	
Identificación de Requisitos	5	17/03/15	19/03/15	
Estimación de Esfuerzo	5	19/03/15	20/03/15	
Plan de Trabajo	5	23/03/15	24/03/15	
Seguimiento del Proyecto Definición	0	24/03/15	24/03/15	Jefe de Proyecto
Diseño Funcional	20	25/03/15	27/03/15	Analista
Análisis y diseño Técnico	**68**	**27/03/15**	**28/04/15**	**Analista**
Análisis y diseño de Modelo de Datos	16	27/03/15	03/04/15	
Análisis y diseño de ETL - Kettle Pentaho	47	03/04/15	24/04/15	
Análisis y diseño de informes - Pentaho Report Designer	5	27/04/15	28/04/15	
Seguimiento del Proyecto Análisis	0	28/04/15	28/04/15	Jefe de Proyecto
Estudio y elaboración de Metadatos	**23**	**28/04/15**	**08/05/15**	**Analista**
Estudio herramenta Pentaho para proyectos PYMES	15	28/04/15	05/05/15	
Construcción del Modelo de Datos	3	05/05/15	06/05/15	
Construcción de Metadatos	5	06/05/15	08/05/15	
Construcción	**159**	**08/05/15**	**22/07/15**	**Programador**
Formación Recursos herramientas Pentaho	30	08/05/15	22/05/15	
Preparación máquina virtual e instalación herramientas	40	22/05/15	10/06/15	
Construcción de ETL - Kettle Pentaho	62	10/06/15	09/07/15	
Construcción de informes - Pentaho Report Designer	27	09/07/15	22/07/15	
Pruebas Unitarias e Integradas	**73**	**22/07/15**	**16/09/15**	**Programador**
Pruebas Unitaras ETL - Kettle Pentaho	47	22/07/15	03/09/15	
Pruebas Integradas ETL - Kettle Pentaho	5	04/09/15	07/09/15	
Pruebas Unitarias informes - Pentaho Report Designer	16	07/09/15	14/09/15	
Pruebas Integradas informes - Pentaho Report Designer	5	15/09/15	16/09/15	
Seguimiento del Proyecto Implantación	0	16/09/15	16/09/15	Jefe de Proyecto
Implantación	**10**	**16/09/15**	**21/09/15**	**Jefe de Proyecto**
Preparar paso a producción	5	16/09/15	18/09/15	
Cumplimiento de Estándares	5	18/09/15	21/09/15	
Formación	10	22/09/15	23/09/15	Analista
Seguimiento Fin del Proyecto	0	23/09/15	23/09/15	Jefe de Proyecto

Tabla 22: Desglose de Tareas

6.4.4 Calendario – Diagrama de Gantt

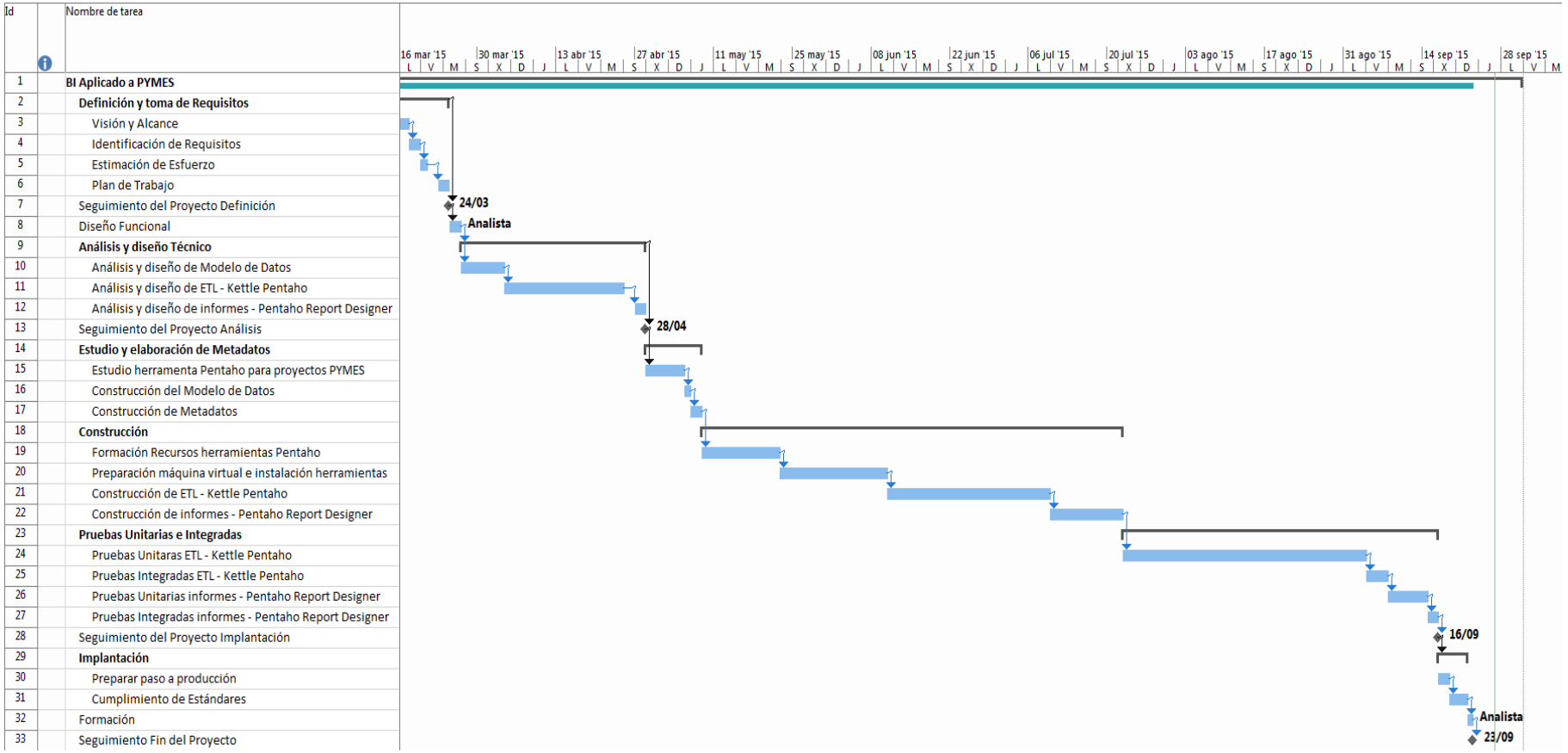


Ilustración 6: Diagrama Gantt

7 DESCRIPCIÓN FUNCIONAL

El objetivo de este documento es describir los aspectos funcionales que debe cumplir el proyecto, que por un lado tiene como objetivo estudiar la viabilidad de desarrollo de una plataforma Business Intelligence orientada a las PYMES, apoyándose para ello en el conjunto de herramientas que ofrece Pentaho en su versión Community, y por otro lado, explotar la información sobre matriculados universitarios que ofrece el Ministerio de Educación, aportando valor añadido.

7.1 PREPARACIÓN DE UNA PLATAFORMA BI

Se debe preparar una plataforma Business Inteligence acondicionada para poder desarrollar cualquier proyecto DWH, apoyándose en las herramientas de Pentaho.

Esta plataforma debe poder alimentarse de cualquier fuente de datos, tanto internos como externos, y estar preparada para soportar el mayor número de gestores de BBDD posible.

7.2 ORIGEN DE DATOS

La fuente de datos de este proyecto serán los ficheros generados con el software PC-AXIS, con extensión *.px, que sube a su web periódicamente el Ministerio de Educación, sobre estadísticas de número de matriculados universitarios[35].

La estadística de Estudiantes Universitarios está recogida en el Plan Estadístico Nacional y aporta información sobre el número de estudiantes matriculados y egresados, así como su género, edad, nacionalidad, lugar de residencia habitual, en el caso de los egresados, tramo de nota del expediente académico, etc.

El marco de la Estadística de Estudiantes Universitarios está definido por todas las universidades, todos sus centros y todas las enseñanzas oficiales que imparten.

7.3 DEFINICIÓN DEL PROCESO DE CARGA

La carga de la información se realizará bajo petición. El Ministerio suele subir las estadísticas cada 6 meses, una con el avance del periodo, y otra con los datos definitivos. Cuando estén subidas las estadísticas definitivas, deberán descargarse en el directorio habilitado para ello, y solicitar al operador que lance el proceso de carga, siguiendo las pautas definidas en el manual de operación.

Se cargará a partir del periodo 2008-2009, ya que es el primer periodo formativo subido en la web del Ministerio.

En cada ejecución se reprocesarán los periodos que estén incluidos en el fichero de metadatos configurado previamente, es decir, podrá cargar un solo periodo, varios e incluso todos, en una sola ejecución.

El proceso deberá ser capaz de actualizar la información, por si hubiera algún error previo que tuviera que corregirse.

7.4 DIMENSIONES DE ANÁLISIS

En este apartado se especifican las dimensiones por las que se puede explotar la información, así como las particularidades que puedan existir en su mantenimiento y proceso de carga.

- Periodo
  El periodo del curso académico es la duración que comprende desde octubre de un año a septiembre del año siguiente, tiempo que incluye todos los periodos de docencia, incluidos los periodos de evaluación.

- Género
  Estudiar el género de los alumnos matriculados puede ser importante ya que la universidad juega un papel crucial en la construcción de la igualdad.

- Ciclo Formativo
  Ciclo en el que está matriculado el alumno:

  - Matrícula de Grado
  - 1º y 2º ciclo. Ciclo Corto
  - 1º y 2º ciclo. Ciclo Largo
  - 1º y 2º ciclo. Solo segundo Ciclo

- Tipo de Universidad
  Tipología de la universidad, pública o privada.

- Universidad o Centro de Enseñanza
  Son los encargados de la organización de las enseñanzas universitarias conducentes a la obtención de títulos académicos.

- Rama
  Son las distintas áreas en que se agrupan los estudios oficiales: Ciencias Sociales y Jurídicas, Ingeniería y Arquitectura, Artes y Humanidades, Ciencias de la Salud y Ciencias.

- Rango de Edad
  Hace referencia a la edad que el estudiante tiene cumplida el 31 de diciembre del primer año de los dos naturales que componen el curso académico.

  Para un mejor análisis, ésta se agrupa en rangos:

  - De 18 a 21 años
  - De 22 a 25 años
  - De 26 a 30 años
  - Más de 30 años

7.5 MEDIDAS

- Número de Alumnos Matriculados
  Número de personas que se encuentran matriculadas, en la fecha de referencia, en una titulación de primer y segundo ciclo, grado, máster o programa de doctorado en alguna universidad española.

- Crecimiento
  El crecimiento es la comparación del número de matriculados en el periodo actual frente al número de matriculados en el periodo anterior. El resultado será un porcentaje, pudiendo ser negativo.

$$\frac{(\text{Nº Matriculados Periodo Actual}) - (\text{Nº Matriculados Periodo Anterior})}{(\text{Nº Matriculados Periodo Anterior})}$$

8 ANÁLISIS Y DISEÑO TÉCNICO

8.1 ANÁLISIS Y DISEÑO DEL MODELO DE DATOS

El modelo de datos que se utiliza en el proyecto es un modelo relacional que sigue un esquema en estrella.

Casi todos los proyectos DWH utilizan el modelo de estrella, o el de copo de nieve, que es una composición de varios modelos de estrella. Se caracteriza por tener una tabla de hechos o agregada que contiene los datos para el análisis, rodeada de tablas de dimensiones.

Estas tablas de dimensiones tienen claves primarias (PK) simples, mientras que las PK de las tablas de hechos normalmente están compuestas por varias PK de diferentes tablas de dimensiones y algún campo que haga referencia al periodo de estudio.

Este tipo de esquema es ideal por su simplicidad y velocidad, ya que permite indexar las dimensiones de forma individualizada sin que repercuta en el rendimiento de la base de datos.

8.1.1 Nomenclatura

Aunque un DWH empiece con muy pocas tablas, y éstas sean fáciles de localizar y manejar, éste tiende a crecer hasta tal punto que puede hacerse muy difícil el mantenimiento del mismo. Para facilitar la tarea, es necesario fijar unas directrices de nomenclatura en los elementos de la base de datos:

- Tablas
    - La primera letra del nombre de la tabla contiene el tipo de tabla. Ésta puede ser **"D"** si es de dimensiones, **"H"**, si es de hechos, **"A"** si es agregada y **"M"** si es de mantenimiento o manual.
    - Los siguientes dos o tres caracteres se utilizan para el código de la iniciativa. En este proyecto se utilizan las letras "EA", de la iniciativa "Estadísticas de Alumnos".
    - El resto de caracteres, hasta un máximo de 15, se utiliza para dar un nombre descriptivo de lo que va a contener la tabla.

Ilustración 7: Nomenclatura

- Campos
  Dependiendo del tipo de campo y/o tipo de datos, el nombre del campo debe comenzar con los siguientes caracteres:

- ID, para los campos clave o identificadores.
- N, para campos numéricos.
- C, para campos de tipo cadena.
- D, para tipo fecha.

A continuación se listan las tablas y campos que conforman el modelo de datos.

8.1.2 Modelo de Datos

8.1.2.1 DEACICLO

Tabla de dimensiones con los diferentes ciclos formativos en los que se puede matricular un alumno. Contiene los siguientes valores:

Matrícula de Grado
1º y 2º ciclo. Ciclo Corto
1º y 2º ciclo. Ciclo Largo
1º y 2º ciclo. Sólo Segundo Ciclo

Columna	Descripción
idciclo	Identificador numérico del ciclo formativo.
cciclo	Descripción del ciclo formativo.
dfeccre	Fecha de carga del ciclo formativo.

Tabla 23: DEACICLO

8.1.2.2 DEAEDAD

Tabla de dimensiones de los distintos rangos de edad de los alumnos.

Columna	Descripción
idrangoedad	Identificador numérico del rango de edad al que pertenece el alumno.
crangoedad	Descripción del rango de edad al que pertenece el alumno.
dfeccre	Fecha de carga del rango de edad.

Tabla 24: DEAEDAD

8.1.2.3 DEAGENERO

Tabla de dimensiones con el género de los alumnos.

Columna	Descripción
idgenero	Identificador numérico del género de los alumnos.
cgenero	Descripción del género de los alumnos.
dfeccre	Fecha de carga del género.

Tabla 25: DEAGENERO

8.1.2.4 DEANACIONALIDAD

Tabla de dimensiones con la nacionalidad de los alumnos.

Columna	Descripción
idnacionalidad	Identificador de la nacionalidad del alumno.
cnacionalidad	Descripción de la nacionalidad del alumno.
dfeccre	Fecha de carga de la nacionalidad.

Tabla 26: DEANACIONALIDAD

8.1.2.5 DEARAMA

Tabla de dimensiones con la rama de estudios de los alumnos.

Columna	Descripción
idrama	Identificador numérico de la rama de estudios.
crama	Descripción de la rama de estudios.
dfeccre	Fecha de carga de la rama de estudios.

Tabla 27: DEARAMA

8.1.2.6 DEATIPUNIV

Tabla de dimensiones con el tipo de universidad. Los valores que toma son “Pública” y “Privada”.

Columna	Descripción
idtipuniv	Identificador numérico del tipo de universidad.
ctipuniv	Descripción del tipo de universidad.
dfeccre	Fecha de carga del tipo de universidad.

Tabla 28: DEATIPUNIV

8.1.2.7 DEAUNIV

Tabla de dimensiones con todas las universidades nacionales, tanto públicas como privadas.

Columna	Descripción
iduniv	Identificador numérico de la universidad.
cuniv	Nombre de la universidad.
idtipuniv	Tipo de universidad, pública o privada. Clave Foránea que cruza con DEATIPUNIV.
dfeccre	Fecha de creación de la universidad.

Tabla 29: DEAUNIV

8.1.2.8 AEAEDAD

Tabla agregada, cuyo indicador es el número de matriculados, y cuyo mínimo nivel de agregación es el rango de edad de los alumnos.

Columna	Descripción
cperiodo	Periodo del ciclo formativo, que comprende desde octubre de un año a septiembre del año siguiente. Contiene los dos años que lo conforman, separados por una barra.
iduniv	Identificador de la universidad. Clave foránea que cruza con DEAUNIV.
idgenero	Identificador del género de los alumnos. Clave foránea que cruza con DEAGENERO.
idciclo	Identificador del ciclo formativo. Clave foránea que cruza con DEACICLO.
idrangoedad	Identificador del rango de edad de los alumnos. Clave foránea que cruza con DEAEDAD.
nnummatriculados	Indicador con el número de alumnos matriculados.

Tabla 30: AEAEDAD

8.1.2.9 AEANACIONALIDAD

Tabla agregada, cuyo indicador es el número de matriculados, y cuyo mínimo nivel de agregación es la nacionalidad de los alumnos.

Columna	Descripción
cperiodo	Periodo del ciclo formativo, que comprende desde octubre de un año a septiembre del año siguiente. Contiene los dos años que lo conforman, separados por una barra.
iduniv	Identificador de la universidad. Clave foránea que cruza con DEAUNIV.
idgenero	Identificador del género de los alumnos. Clave foránea que cruza con DEAGENERO.
idciclo	Identificador del ciclo formativo. Clave foránea que cruza con DEACICLO.
idnacionalidad	Identificador de la nacionalidad de los alumnos. Clave foránea que cruza con DEANACIONALIDAD.
nnummatriculados	Indicador con el número de alumnos matriculados.

Tabla 31: AEANACIONALIDAD

8.1.2.10 AEARAMA

Tabla agregada, cuyo indicador es el número de matriculados, y cuyo mínimo nivel de agregación es la rama de estudios de los alumnos.

Columna	Descripción
cperiodo	Periodo del ciclo formativo, que comprende desde octubre de un año a septiembre del año siguiente. Contiene los dos años que lo conforman, separados por una barra.
iduniv	Identificador de la universidad. Clave foránea que cruza con DEAUNIV.
idgenero	Identificador del género de los alumnos. Clave foránea que cruza con DEAGENERO.
idciclo	Identificador del ciclo formativo. Clave foránea que cruza con DEACICLO.
idrama	Identificador de la rama de estudios de los alumnos. Clave foránea que cruza con DEARAMA.
nnummatriculados	Indicador con el número de alumnos matriculados.

Tabla 32: AEARAMA

Puede parecer que las tablas agregadas del modelo tienen información redundante, pero hay que tener en cuenta que el origen de datos ya trae la información agregada, por lo que ésta es la única forma de poder analizar toda la información de origen posible.

En el siguiente punto se detallan los modelos Entidad-Relación del proyecto, generados con la herramienta Microsoft Visio 2013.

8.1.3 Modelo Rango Edad

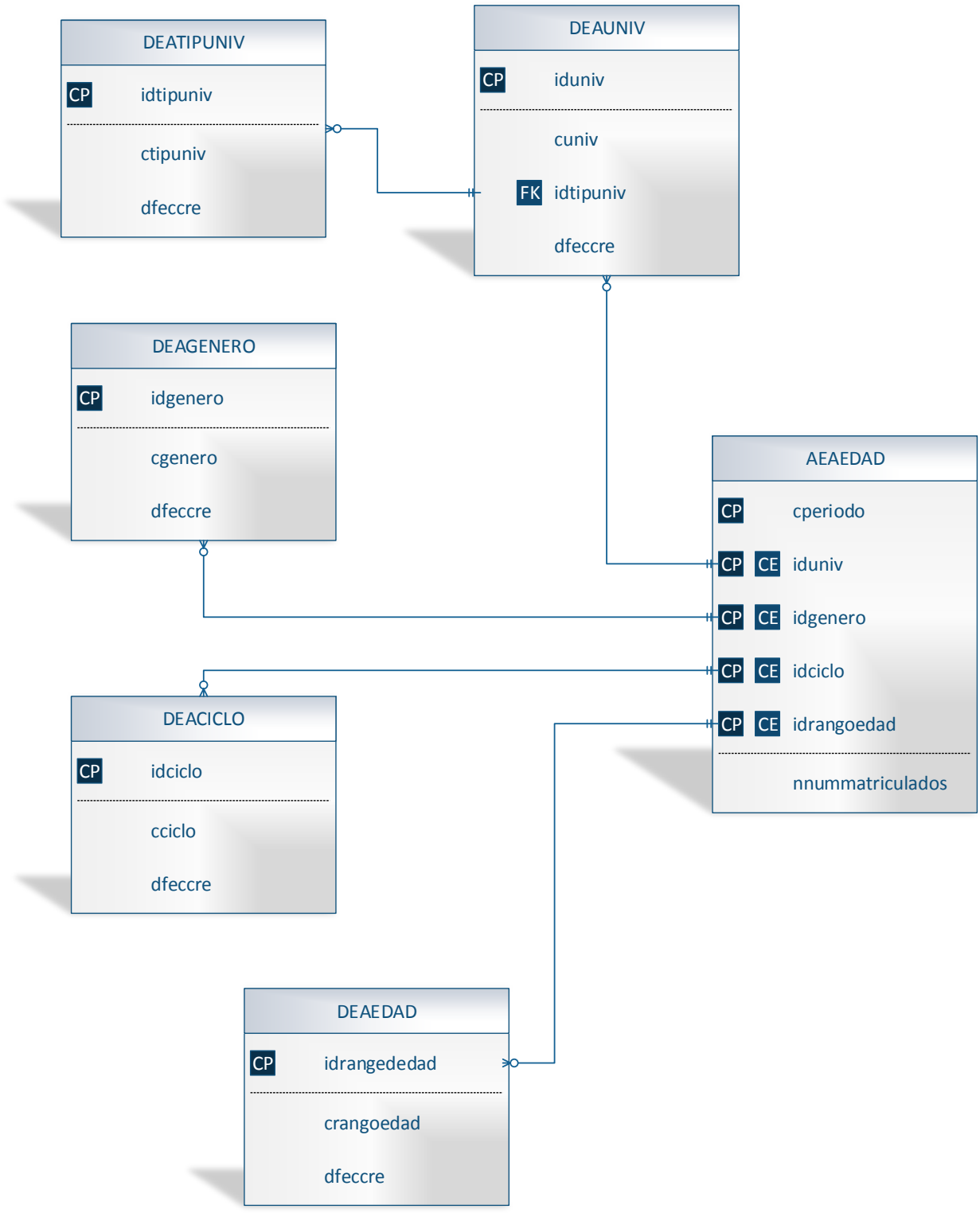


Ilustración 8: Modelo Rango Edad

8.1.4 Modelo Nacionalidad

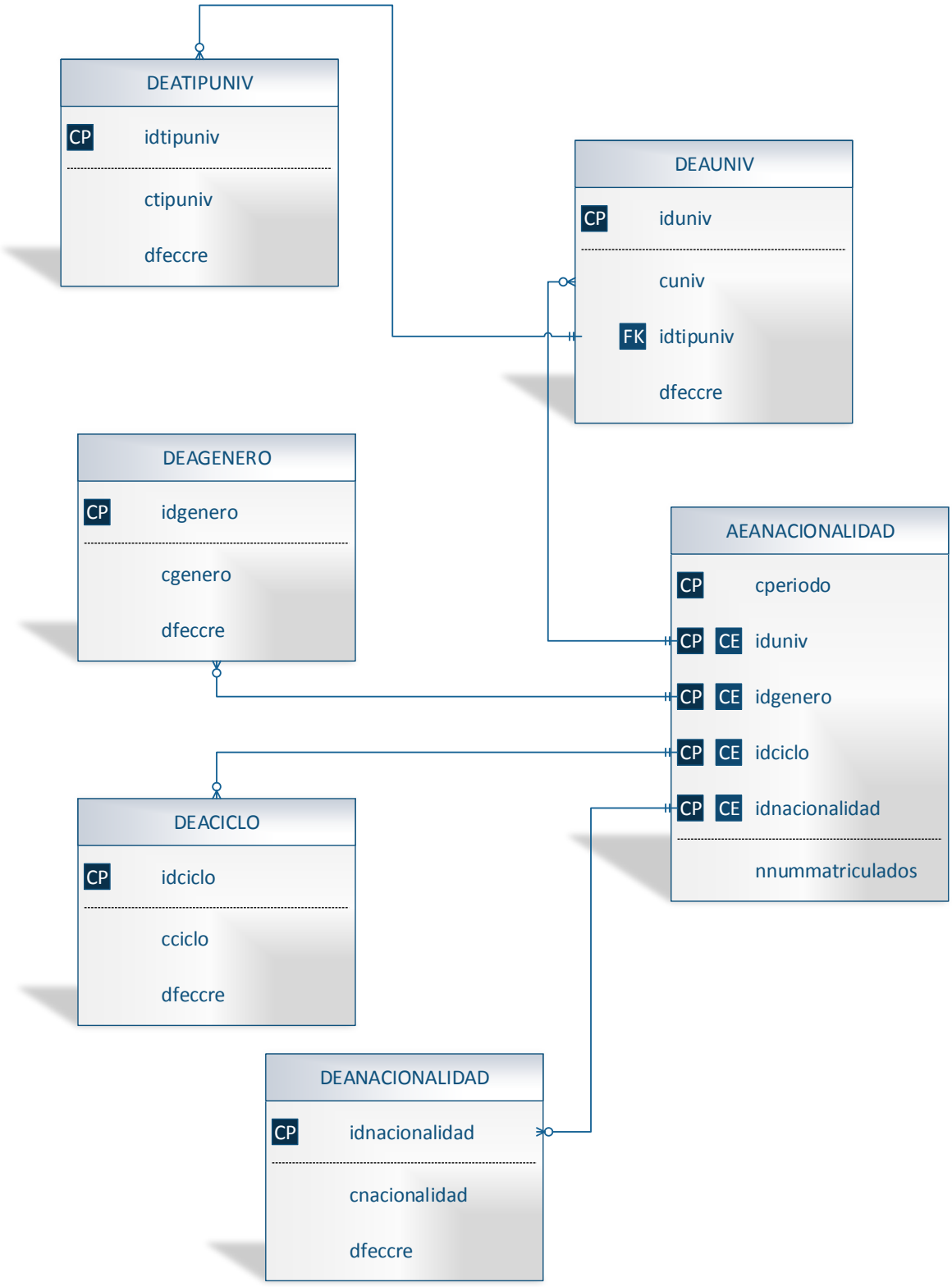


Ilustración 9: Modelo Nacionalidad

8.1.5 Modelo Rama

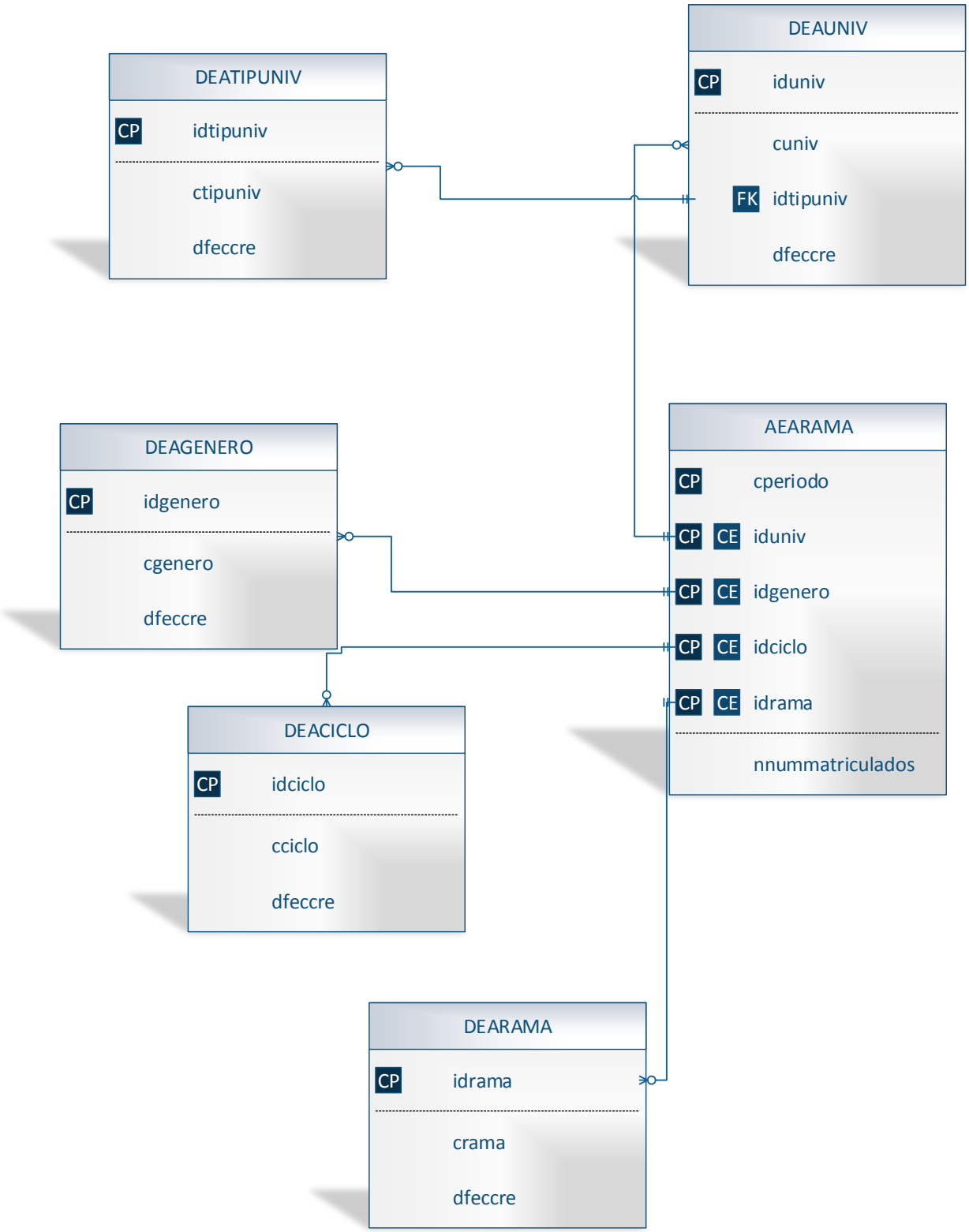


Ilustración 10: Modelo Rama

8.2 DISEÑO TÉCNICO ETL

Los procesos de Extracción, Tratamiento y Carga de este proyecto se desarrollan con la herramienta Kettle de Pentaho, ayudados de su interfaz gráfica integrada llamada Spoon.

Con Kettle se pueden definir dos tipos de procesos, transformaciones y trabajos. La diferencia entre ambos es que mientras las transformaciones se utilizan para mover y transformar filas entre una fuente y un destino, los trabajos permiten controlar el flujo a más alto nivel: ejecutar transformaciones, enviar un correo electrónico en caso de error, enviar archivos por FTP, etc.

Por otro lado, se definen los scripts de Perl para la transformación de archivos PC-AXIS en ficheros planos separados por tabulaciones. Estos scripts deben ir integrados en la herramienta Spoon.

8.2.1 Script PC-Axis a relacional

Este script ha de desarrollarse en Perl, apoyándose en el módulo **Data::PcAxis** incluido en el CPAN de Perl, desarrollado por Fiachra O'Donoghue[14].

Este módulo permite leer ficheros con el formato PC-AXIS, almacenándolos en arrays multidimensionales.

Si se abren con un editor de texto, los archivos PC-AXIS siguen un formato como el del siguiente ejemplo:

AXIS-VERSION="2000";
DECIMALS=0;
SHOWDECIMALS=0;
MATRIX="AV1603f";
SUBJECT-CODE="AV1";
SUBJECT-AREA="Avance de la Estadística de Estudiantes Universitarios. Curso 2014/2015";
COPYRIGHT=YES;
DESCRIPTION="III.2.3f Número total de estudiantes matriculados, por grupo de edad, sexo y universidad. Estudios de 1º y 2º Ciclo: Solo Segundo Ciclo.";
TITLE="III.2.3f Número total de estudiantes matriculados, por grupo de edad, sexo y universidad. Estudios de 1º y 2º Ciclo: Solo Segundo Ciclo.";
DESCRIPTIONDEFAULT=YES;
CONTENTS="Estudiantes matriculados";
UNITS="Estudiantes matriculados";
STUB="Universidad";
HEADING="Grupo de Edad","Sexo";
VALUES("Universidad")="Total Universidades Privadas","Antonio de Nebrija",
"Católica de Valencia San Vicente Mártir","Católica San Antonio","Deusto","Europea de Madrid",
"Mondragón Unibertsitatea","Navarra","Pontificia Comillas","Ramón Llull","San Pablo-CEU",
"Vic-Central de Catalunya";
VALUES("Grupo de Edad")="Todas las edades","De 22 a 25 años","De 26 a 30 años","Más de 30 años";
VALUES("Sexo")="Ambos Sexos","Mujeres","Hombres";
SOURCE="S.G. de Coordinación y Seguimiento Universitario. Ministerio de Educación, Cultura y Deporte.";
NOTEX="Todas las universidades son presenciales.";
DATA=
167 61 106 18 11 7 34 13 21 115 37 78
1 "." 1 "." "." "." 1 "." 1 "." "." "."

2 2 "." "." "." "." "." "." "." "." 2 2 "."
11 5 6 "." "." "." 1 1 "." 10 4 6
9 4 5 2 "." 2 2 "." 2 5 4 1
21 7 14 3 3 "." 4 "." 4 14 4 10
10 8 2 2 2 "." 5 4 1 3 2 1
5 "." 5 3 "." 3 1 "." 1 1 "." 1
6 1 5 3 1 2 3 "." 3 "." "." "."
7 "." 7 "." "." "." "." "." "." "." 7 "." 7
7 7 "." 5 5 "." 1 1 "." 1 1 "."
88 27 61 "." "." "." 16 7 9 72 20 52 ;

Las primeras líneas contienen metadatos, como el nombre de la estadística, descripción, el indicador objeto de la estadística, etc. Aunque los campos que realmente importan son "VALUES", ya que contienen los valores de cada una de las dimensiones, y el campo "DATA", que contiene los valores de los indicadores.

El objetivo del script es obtener un fichero con un formato como el siguiente:

A Coruña Todas las edades Ambos Sexos 13345
A Coruña Todas las edades Mujeres 6842
A Coruña Todas las edades Hombres 6503
A Coruña De 18 a 21 años Ambos Sexos 7180
A Coruña De 18 a 21 años Mujeres 3935
A Coruña De 18 a 21 años Hombres 3245
A Coruña De 22 a 25 años Ambos Sexos 4198
A Coruña De 22 a 25 años Mujeres 2001
A Coruña De 22 a 25 años Hombres 2197
A Coruña De 26 a 30 años Ambos Sexos 1124
A Coruña De 26 a 30 años Mujeres 496
A Coruña De 26 a 30 años Hombres 628
A Coruña Más de 30 años Ambos Sexos 843
A Coruña Más de 30 años Mujeres 410
A Coruña Más de 30 años Hombres 433
Alcalá Todas las edades Ambos Sexos 13600
Alcalá Todas las edades Mujeres 7666
Alcalá Todas las edades Hombres 5934
…

Además, debe permitir incluir información adicional que no esté de forma implícita en el fichero PC-AXIS. Para ello, el script debe apoyarse en un fichero de metadatos, llamado ff_metadatos.txt que siga el siguiente formato:

Campo Opcional 1	Campo Opcional 2	Campo Opcional 3	Campo Opcional 4	Nombre fichero Entrada	Nombre fichero salida	Charset fichero entrada

Tabla 33: Formato Fichero Metadatos

Los cuatro primeros campos son opcionales. El valor que contengan será fijo para todas las filas de la estadística del fichero PC-AXIS. Por ejemplo, si se está cargando un fichero con las estadísticas de los

alumnos matriculados en el ciclo corto, pero este valor no está incluido en el contenido de la estadística, ese valor se puede incluir aquí para no perder esa información.

Un ejemplo de fichero de metadatos sería el siguiente:

2013/2014 Matrícula de Grado Pública 20132014_III3b_pub.px ff_edad.txt cp437
2013/2014 1º y 2º ciclo. Ciclo Corto Pública 20132014_III3d_pub.px ff_edad.txt cp437
2013/2014 1º y 2º ciclo. Ciclo Largo Pública 20132014_III3e_pub.px ff_edad.txt cp437
2013/2014 1º y 2º ciclo. Solo segundo Ciclo Pública 20132014_III3f_pub.px ff_edad.txt cp437
…

El resultado final, incluyendo la información del fichero de metadatos, quedaría así:

2013/2014 Matrícula de Grado Pública A Coruña Todas las edades Ambos Sexos 13345
2013/2014 Matrícula de Grado Pública A Coruña Todas las edades Mujeres 6842
2013/2014 Matrícula de Grado Pública A Coruña Todas las edades Hombres 6503
2013/2014 Matrícula de Grado Pública A Coruña De 18 a 21 años Ambos Sexos 7180
2013/2014 Matrícula de Grado Pública A Coruña De 18 a 21 años Mujeres 3935
…

Los ficheros resultantes de este scripts serán los que alimentarán las transformaciones desarrolladas en Kettle de Pentaho.

Este script debe estar preparado para cualquier fichero PC-AXIS, no solo para las estadísticas de alumnos universitarios del Ministerio de Educación.
Además, debe estar acompañado de un launcher que lo ejecute para todos los ficheros PC-AXIS que estén incluidos en el fichero de metadatos.

Para el desarrollo, se debe partir de una función que utilice recursividad, para poder recorrer todos los nodos del array multidimensional en el que se almacena la estadística, mientras se va escribiendo en el fichero de salida.

8.2.2 Transformaciones y trabajos de Kettle Pentaho

Para todas las transformaciones y trabajos de Kettle de este proyecto, se utiliza la siguiente conexión de BBDD, aunque la herramienta permite tantas conexiones como sean necesarias, de diferentes gestores de base de datos.

Para crearla es necesario ir al panel izquierdo de Kettle, en la pestaña “view”, dentro de “Transformations”. En el menú contextual de “Database conection”, se pincha en new.

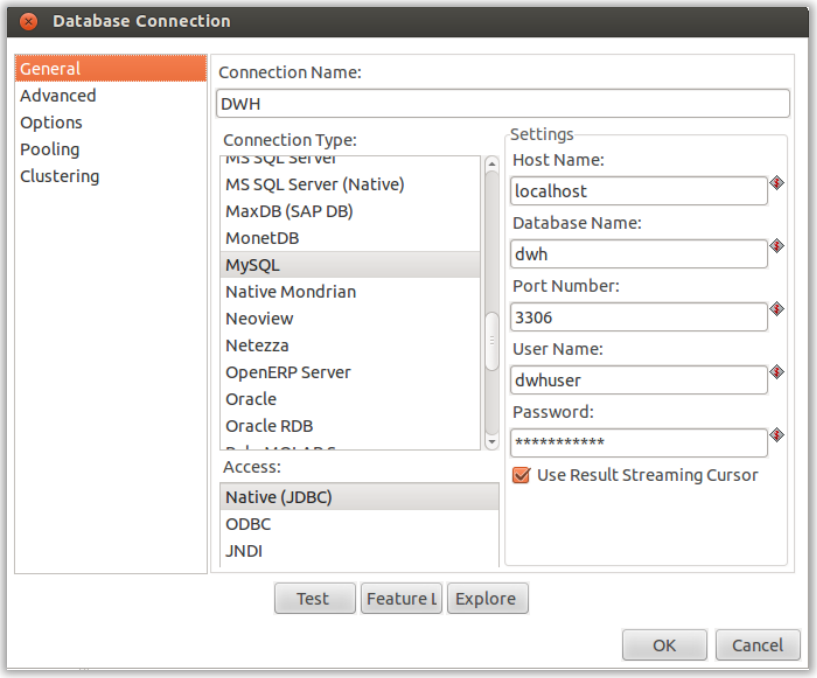


Ilustración 11: Conexión BBDD Kettle

Se utilizan los datos de la BBDD "dwh" creada para el proyecto. El nombre de la conexión es también "DWH".

Nombre de la Base de datos: dwh
Puerto: 3306
Nombre de usuario: dwhuser
Contraseña: *pentaho2014*

Las variables utilizadas, comunes para todos los procesos de Kettle del proyecto, son:

Directorio con los scripts de perl
v_dirscripts = /home/pentaho/pentaho-files/scripts/perl/

Directorio con los ficheros de entrada
v_dirsrcfiles = /home/pentaho/pentaho-files/srcfiles/

8.2.3 t_dimensiones_edad

Transformación que parte de un fichero resultante del script descrito en el anterior punto, habiendo sido llamado ff_edad.txt, y que carga todos los valores posibles de todas las dimensiones del modelo de edad. Estas tablas son:

- DEACICLO
- DEAGENERO
- DEAEDAD
- DEAUNIV
- DEATIPUNIV

La transformación sigue el siguiente flujo:

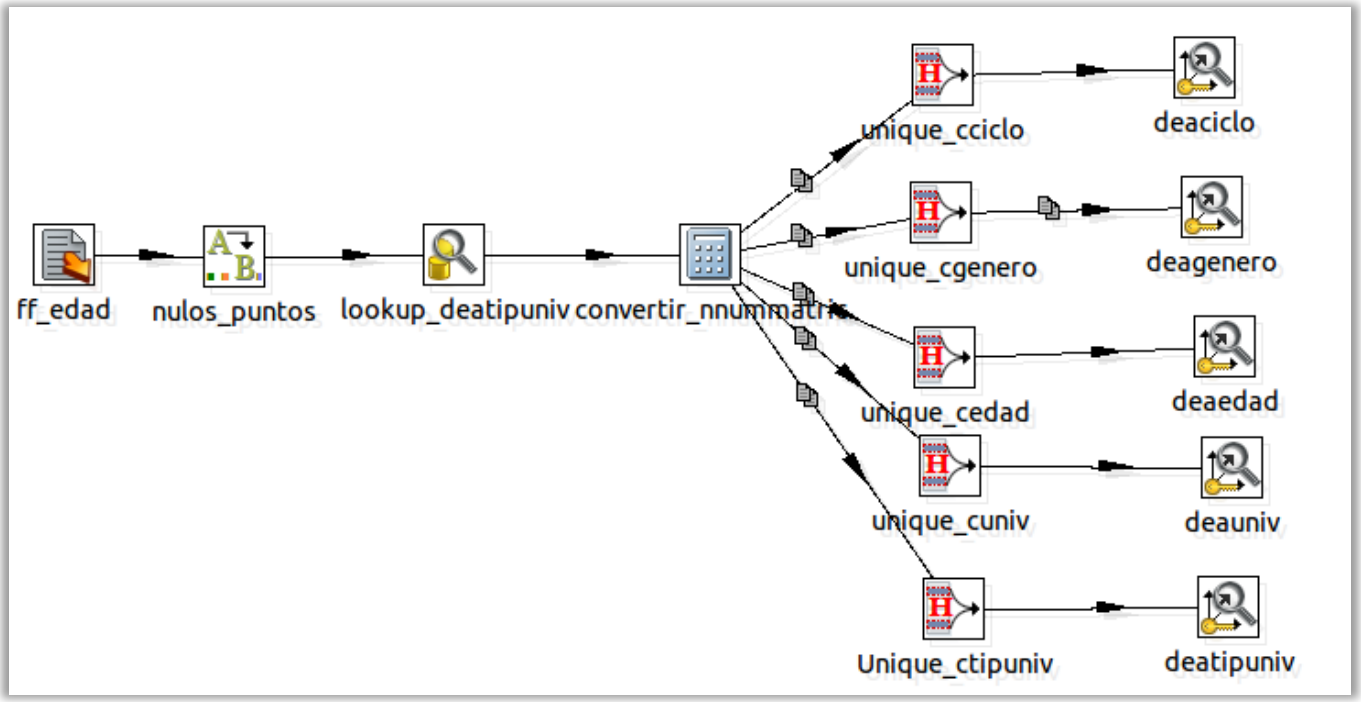


Ilustración 12: Transformación t_dimensiones_edad

A continuación se describen cada uno de los objetos que forman la transformación t_dimensiones_edad.

8.2.3.1 ff_edad

Objeto del tipo “Text file input”, cuya función es leer un fichero de texto de entrada. Debe definirse con los siguientes valores:

- Directorio y nombre del fichero. El nombre del fichero es “ff_edad.txt”, y su ubicación está definida en la variable “v_dirsrcfiles”. La llamada a las variables en Kettle se realiza utilizando los caracteres ${*nombre_variable*}.

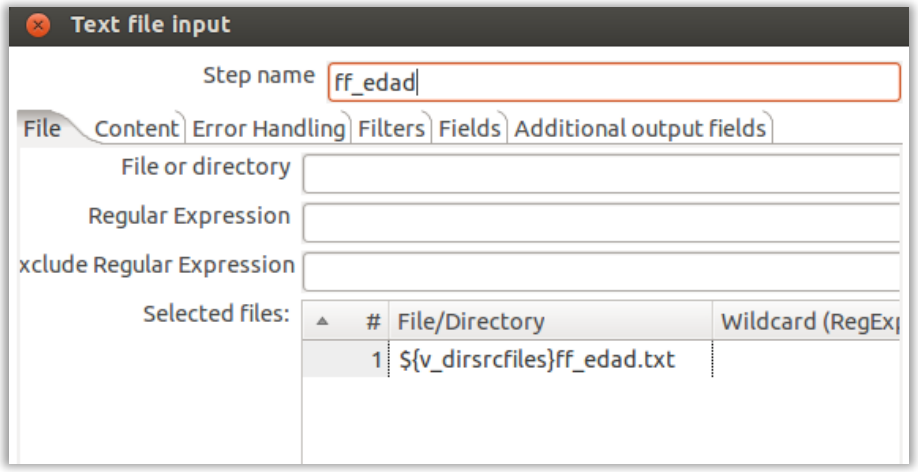


Ilustración 13: Objeto Text File input – Pestaña file

- Contenido. Fichero de texto cuyos campos están separados por tabulaciones, por lo que se define como “CSV”. Además, se marca la opción de “No empty rows” para ignorar las filas vacías.

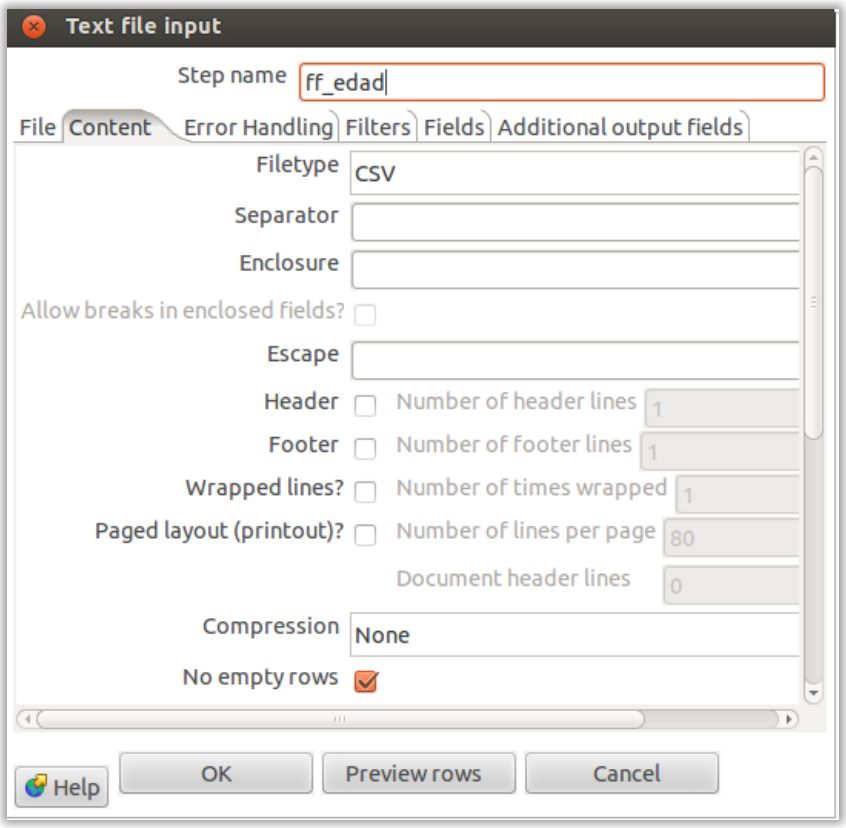


Ilustración 14: Objeto Text File input – Pestaña Content

- Filtros. Las estadísticas contienen filas de resumen con totales que no interesan cargar. Éstas contienen los literales "total", "Todas", "Ambos", "Total" y "Presenciales". Se incluyen en la pestaña de filtros para que el proceso las descarte.

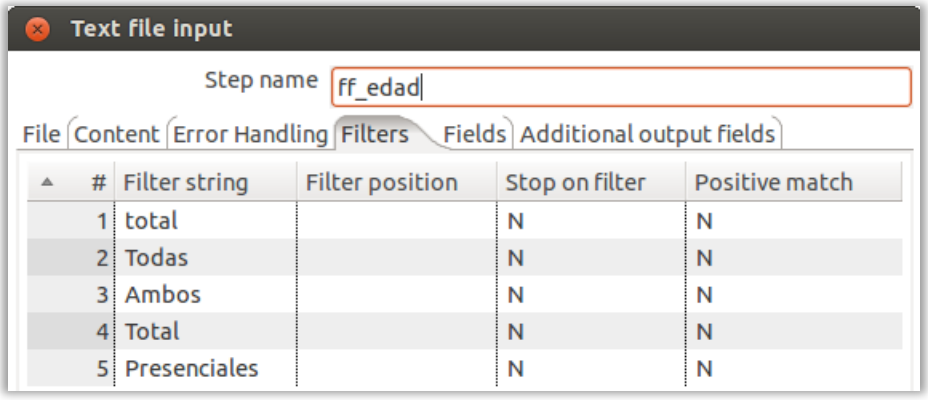
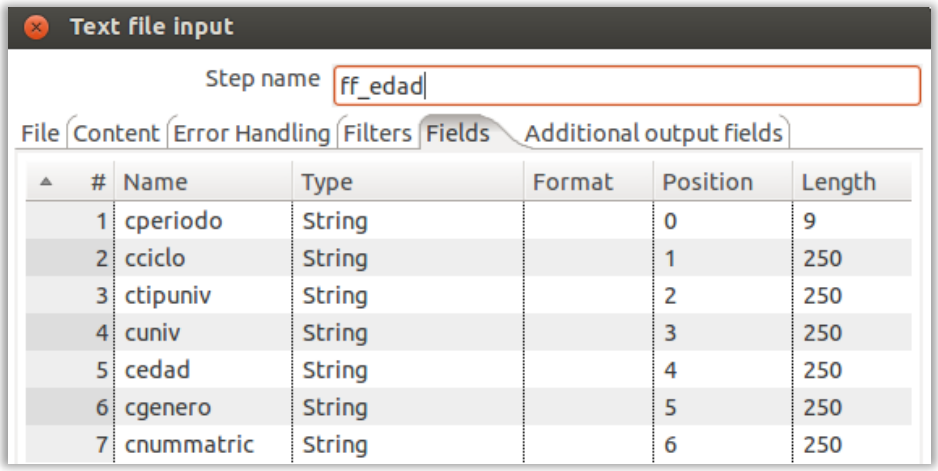


Ilustración 15: Objeto Text File input – Pestaña Filters

- Campos. Se listan los campos que forman el fichero, con su tipo de datos, tamaño y posición.



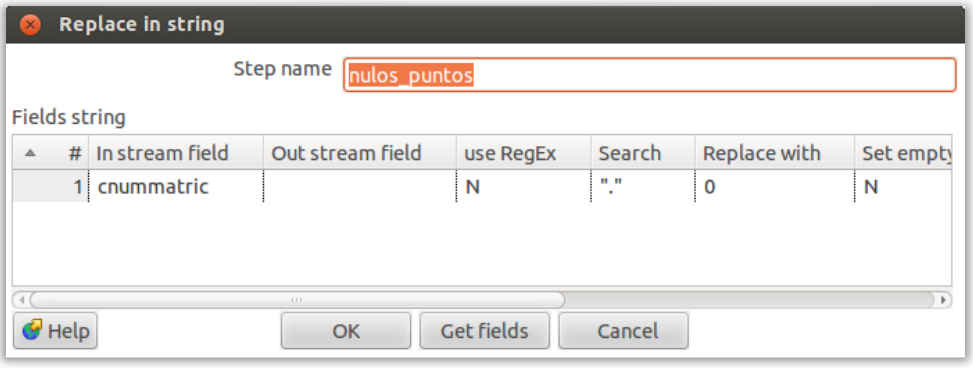
#	Name	Type	Format	Position	Length
1	cperiodo	String		0	9
2	cciclo	String		1	250
3	ctipuniv	String		2	250
4	cuniv	String		3	250
5	cedad	String		4	250
6	cgenero	String		5	250
7	cnummatric	String		6	250

Ilustración 16: Objeto Text File input – Pestaña Fields

Para asegurarse de que se ha definido correctamente el objeto "Text file input", y en general, para todos los objetos, Kettle incluye la opción de previsualizar filas, que genera una muestra de lo que se está desarrollando.

8.2.3.2 nulos_puntos

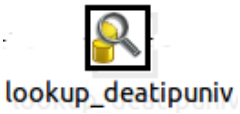
Objeto del tipo "Replace in string" cuya función es remplazar un carácter o un conjunto de caracteres por otro/s, para un campo determinado. En este caso se pretende eliminar el punto de separación de miles del campo cnummatric.



#	In stream field	Out stream field	use RegEx	Search	Replace with	Set empty
1	cnummatric		N	"."	0	N

Ilustración 17: Objeto Replace in string

8.2.3.3 lookup_deatipuniv



Objeto del tipo “Database Lookup” cuya función es buscar un valor en una tabla determinada, normalmente de dimensiones, para devolver otro, como si de una función se tratara. En este caso se pretende buscar el identificador asociado a la descripción del tipo de universidad (pública o privada), en la tabla DEATIPUNIV, para posteriormente cargarlo en la tabla DEAUNIV. Si el lookup no devolviera ningún valor, se le asignaría -1.

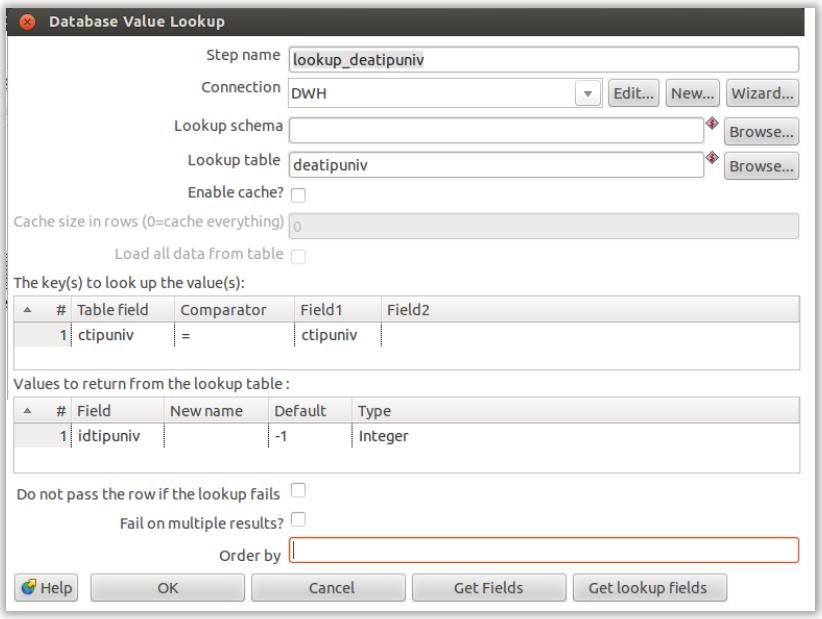


Ilustración 18: Objeto Database Lookup

8.2.3.4 convertir_nnummatric

El objeto “Calculator”, puede realizar un gran abanico de operaciones, tales como operaciones matemáticas, sobre fechas, cadenas, etc. En este caso el objetivo es realizar una copia del campo cnummatric, aprovechando que se puede definir el tipo de dato destino, para convertirlo en numérico. El nuevo campo se llama nnummatric.

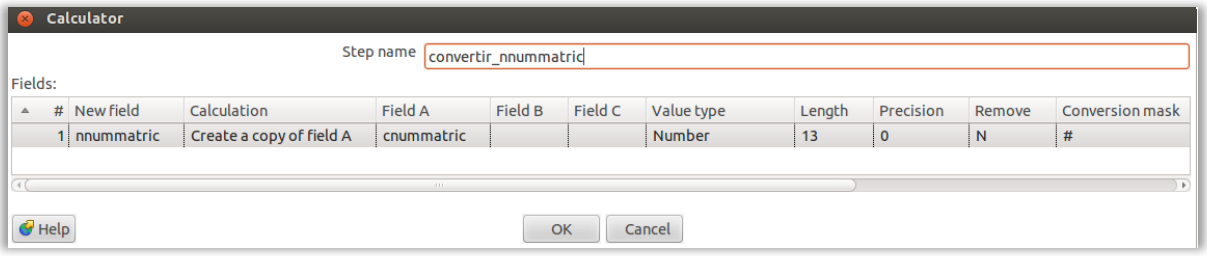


Ilustración 19: Objeto Calculator

8.2.3.5 unique_cciclo

Objeto del tipo "Unique rows (Hashset)" que funciona como un distinct para los campos seleccionados. En este caso, el distinct se realiza sobre cciclo, aunque en la transformación hay un objeto de este tipo por dimensión a cargar: cgenero, cedad, cuniv y ctipuniv.

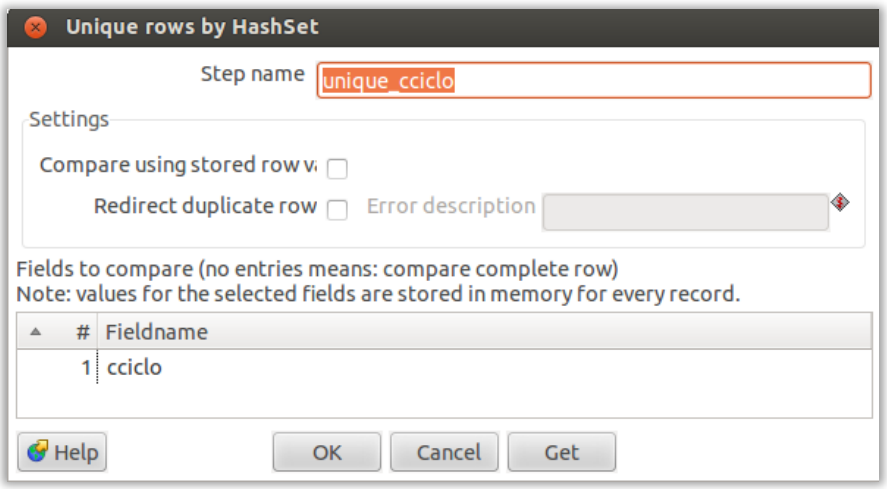


Ilustración 20: Objeto Unique rows (Hashet)

8.2.3.6 deaciclo

Objeto "Combination Lookup / Update", preparado para las dimensiones del tipo 1 según la clasificación de Kimball de dimensiones lentamente cambiantes.

Las dimensiones de tipo 1 son las más básicas y sencillas de implementar, ya que si bien no guardan los cambios históricos, tampoco requieren ningún modelado especial y no necesitan que se añadan nuevos registros a la tabla.

Cuando un registro presente un cambio en alguno de los valores de sus campos, se debe proceder simplemente a actualizar el dato en cuestión, sobrescribiendo el antiguo.

En este caso, la dimensión a rellenar es DEACICLO, aunque la carga del resto de dimensiones sigue el mismo funcionamiento.

El campo identificador se calcula usando el máximo de la tabla más uno.

Además, se rellena el campo dfeccre con la fecha en el que se inserta el dato.

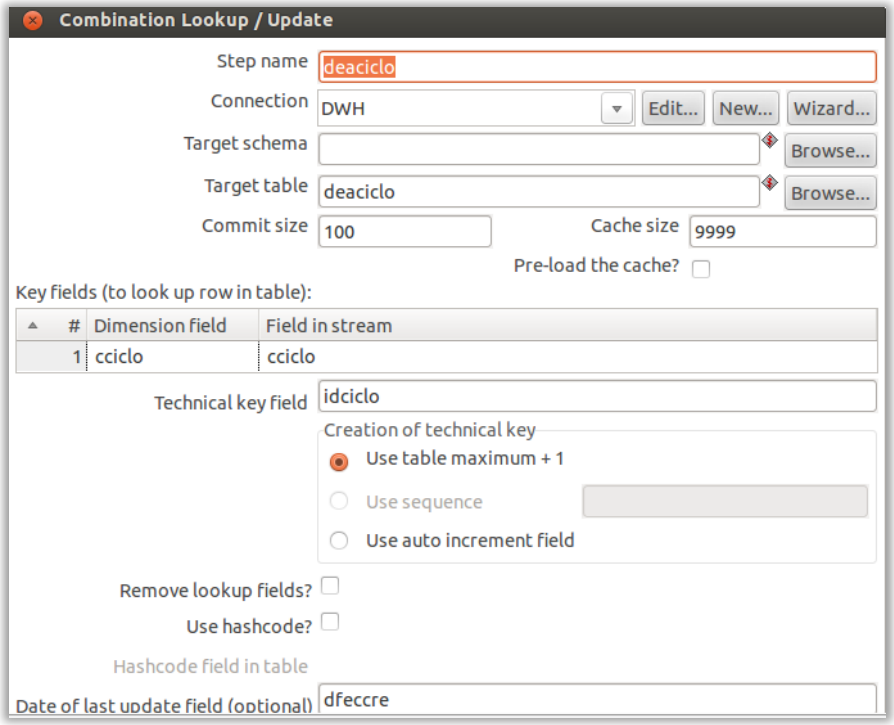


Ilustración 21: Objeto Combination Lookup/Update

8.2.4 t_dimensiones_rama

Esta transformación es idéntica a "t_dimensiones_edad", salvo que el fichero de origen es "ff_rama.txt", estadística que tiene la dimensión rama en lugar de la dimensión edad. Por lo que también, en lugar de cargar DEAEDAD, carga DEARAMA:

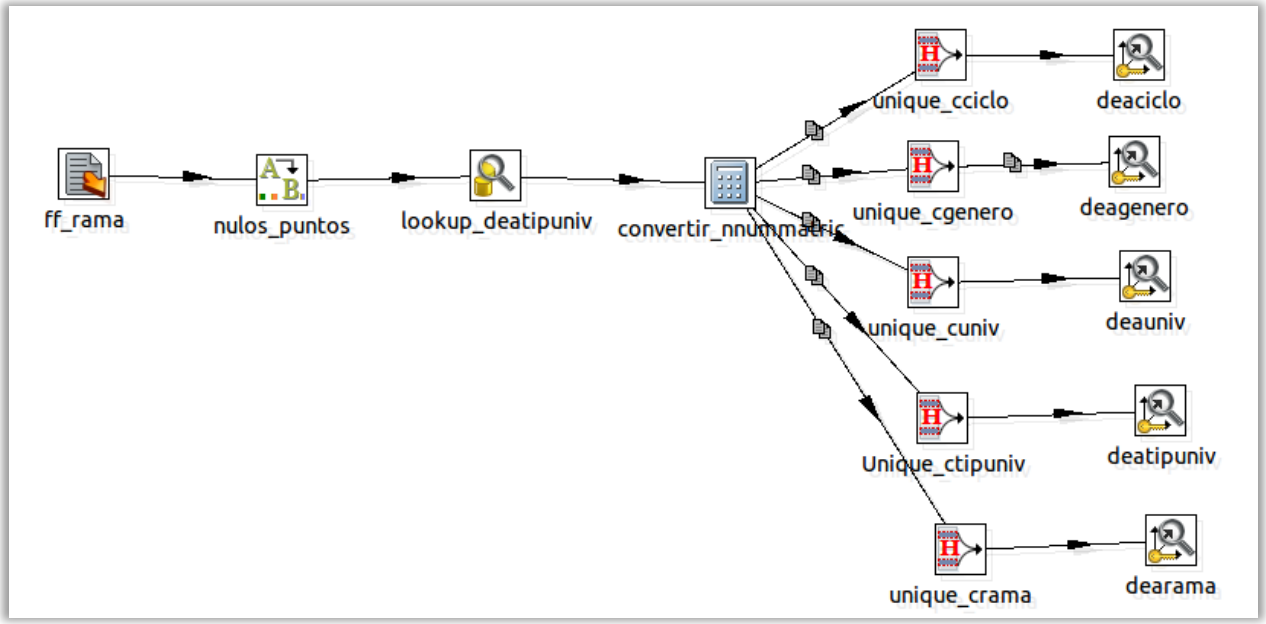


Ilustración 22: Transformación t_dimensiones_rama

8.2.5 t_agregada_edad

Transformación cuyo objetivo es cargar la tabla AEAEDAD a partir del fichero de texto separado por tabulaciones "ff_edad.txt". El proceso actualiza, inserta o elimina según convenga, evaluando solamente el periodo de tiempo, y el ciclo formativo que venga definido en el fichero de entrada. Es decir, si el fichero sólo trae la estadística del curso académico 2014/2015, y en la tabla estaban ya cargados todos los periodos anteriores, el proceso no debe borrar ningún registro anterior a 2014/2015.

Tiene 3 flujos que terminan divergiendo en uno solo, que será el que inserte, elimine o actualice en AEAEDAD:

- Flujo Source: Es el flujo de datos que lee a partir del fichero ff_edad.txt, y que prepara los campos para adecuarlos a la tabla AEAEDAD, buscando en sus tablas de dimensiones, el identificador correspondiente en cada caso.
- Flujo Comprobación INS/UPD/DEL: Flujo que lee de la tabla destino, AEAEDAD, para posteriormente cruzar con el Flujo Source, y comprobar que los registros que vienen de éste último se tienen que insertar, actualizar, eliminar, o si por el contrario no se debe hacer nada.
- Flujo seleccionar periodo a evaluar y cargar: Flujo que lee también del fichero ff_edad.txt, pero en este caso, simplemente para seleccionar los periodos y ciclos formativos a tener en cuenta, para posteriormente cruzar con el flujo anterior.

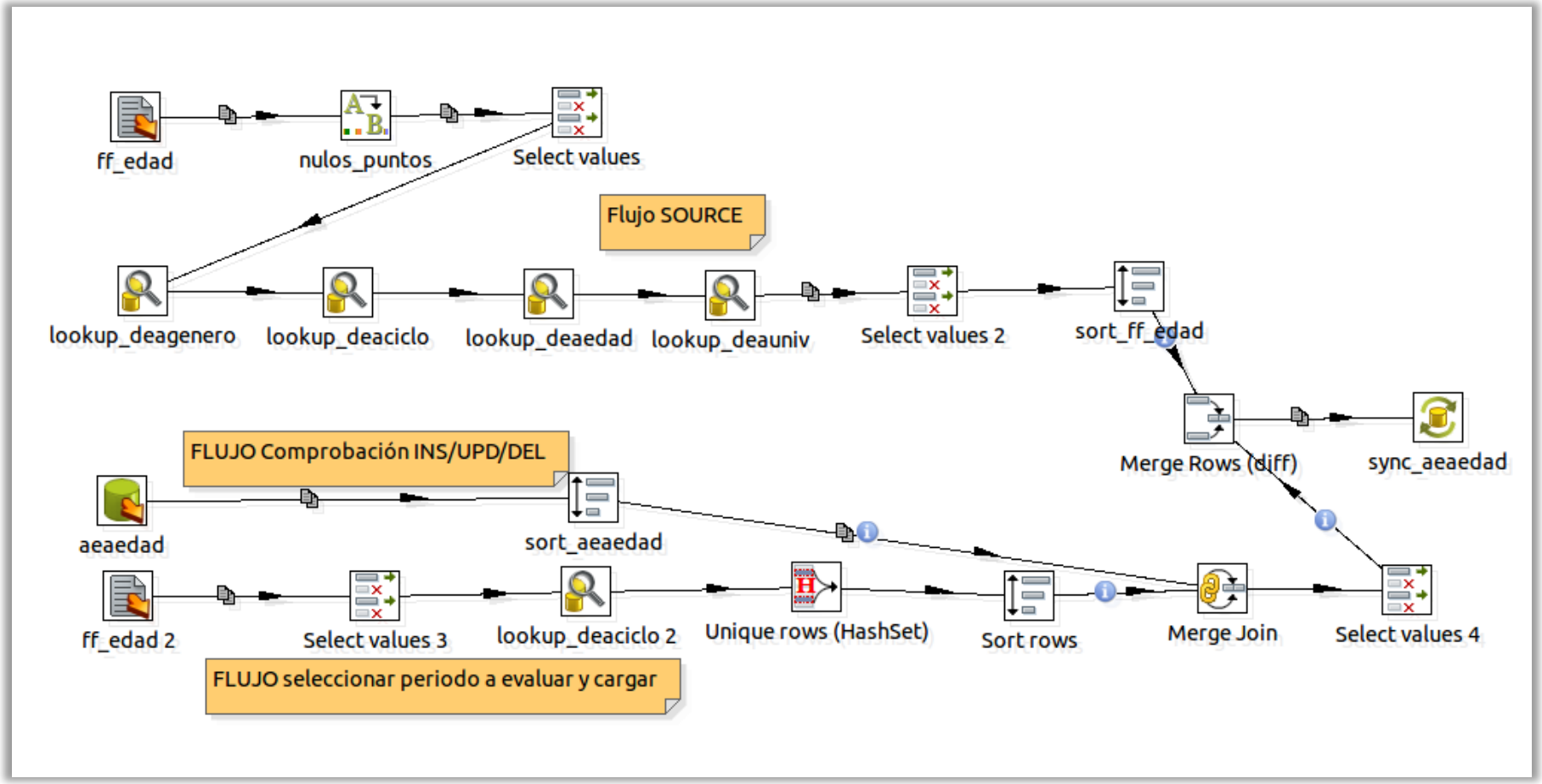


Ilustración 23: Transformación t_agregada_edad

Objetos de la transformación t_agregada_edad:

Los objetos ff_edad, ff_edad 2 y nulos_puntos son exactamente igual que los de la transformación t_dimensiones_edad.

8.2.5.1 Select_values

Objeto del tipo “Select Values”, que permite seleccionar, eliminar o renombrar campos del flujo de datos. En este caso se busca renombrar el campo cnummatric por nnummatriculados, y aprovechar para cambiar el tipo de datos del mismo, de texto a numérico. Esto se puede hacer en la pestaña “meta-data” del objeto.

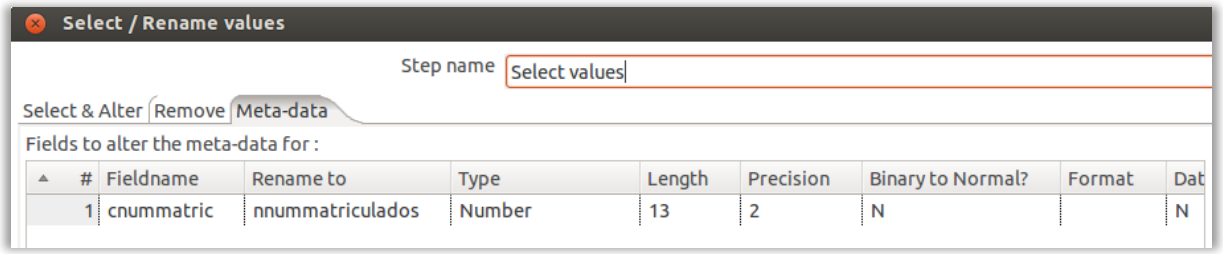
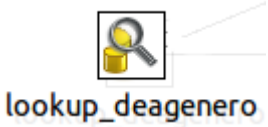


Ilustración 24: Objeto Select Values – Meta-data

8.2.5.2 lookup_deagenero



Objeto tipo “Database Lookup” cuya función es buscar el correspondiente identificador idgenero, en la tabla DEAGENERO, a partir de su descripción cgenero. Este tipo de objeto ya se ha explicado en la transformación t_dimensiones_edad, y el funcionamiento en este caso es el mismo. Se puede decir lo mismo de los objetos lookup_deaciclo, lookup_deaciclo 2, lookup_deaedad, lookup_deainiv.

8.2.5.3 Select Values 2

En este caso, el objeto del tipo “Select Values” se utiliza para seleccionar solamente los campos necesarios para cargar AEAEDAD, prescindiendo del resto. Esto se define en la pestaña “Select & Alter”.

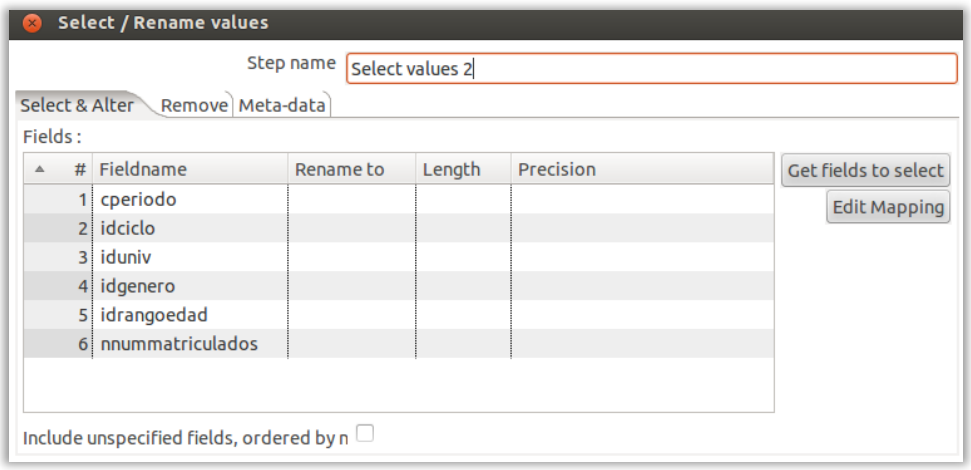


Ilustración 25: Objeto Select Values – Select&Alter

8.2.5.4 sort_ff_edad

Objeto del tipo "Sort rows", que permite ordenar por los campos especificados. En este caso, se desea ordenar por cperiodo, idciclo, iduniv, idgenero e idrangoedad, de forma ascendente, en ese orden.

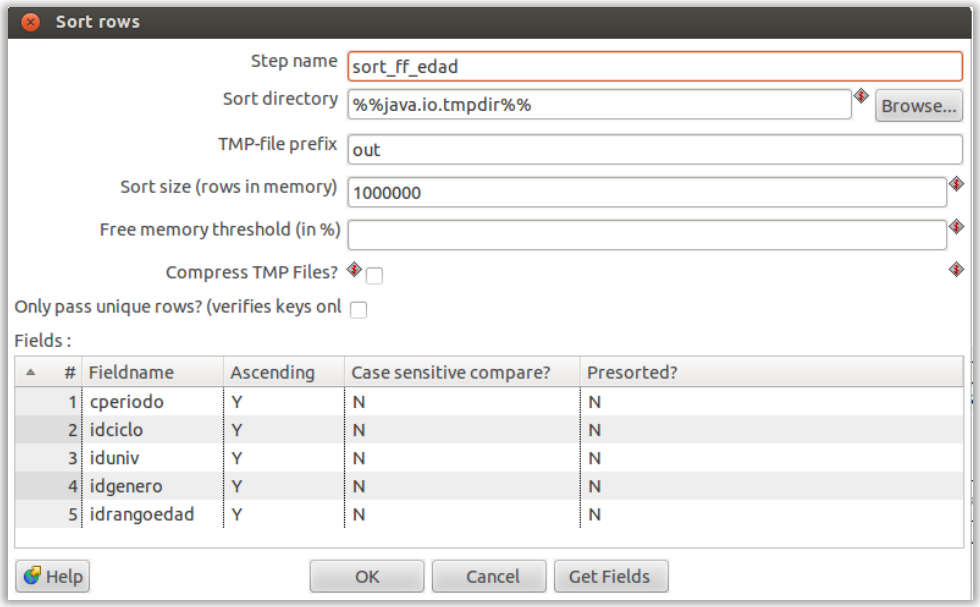


Ilustración 26: Objeto Sort Rows

8.2.5.5 aeaedad

Objeto tipo "Table input" que realiza una lectura de una tabla de una base de datos. En este caso, la lectura se hace de la tabla AEAEDAD, utilizando la conexión "DWH".

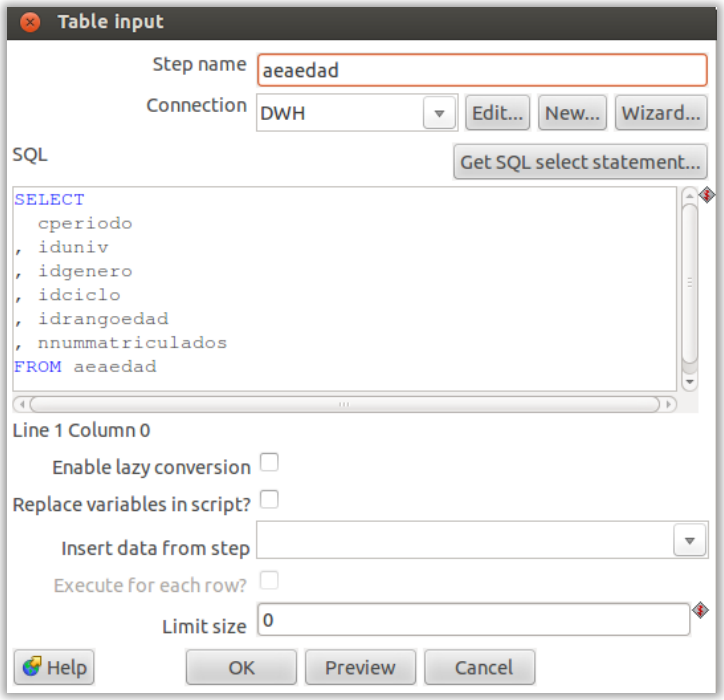


Ilustración 27: Objeto Table input

8.2.5.6 sort_aeaedad

Objeto "Sort rows", utilizado para ordenar los campos cperiodo, idciclo, iduniv, idgenero e idrangoedad, de forma ascendente.

8.2.5.7 Select Values 3

Objeto "Select Values" cuya función es seleccionar los campos cperiodo y cciclo.

8.2.5.8 Unique Rows (HashSet)

Realiza el distinct de los campos cperiodo y cciclo.

8.2.5.9 Sort rows

Ordena de forma ascendente cperiodo y cciclo.

8.2.5.10 Merge Join

Objeto tipo “Merge Join” que realiza el cruce entre dos flujos, pudiendo elegir entre diferentes estrategias: INNER, LEFT OUTER, RIGTH OUTER o FULL OUTER.

El objetivo es obtener todo lo que exista en AEAEDAD para los periodos y ciclos que existan en el fichero ff_edad.txt, por lo que la estrategia a seguir es un INNER JOIN.

En este tipo de objetos, es necesario definir cuál es el flujo primario y cuál es el secundario, aunque en este caso es irrelevante, al ser un INNER JOIN.

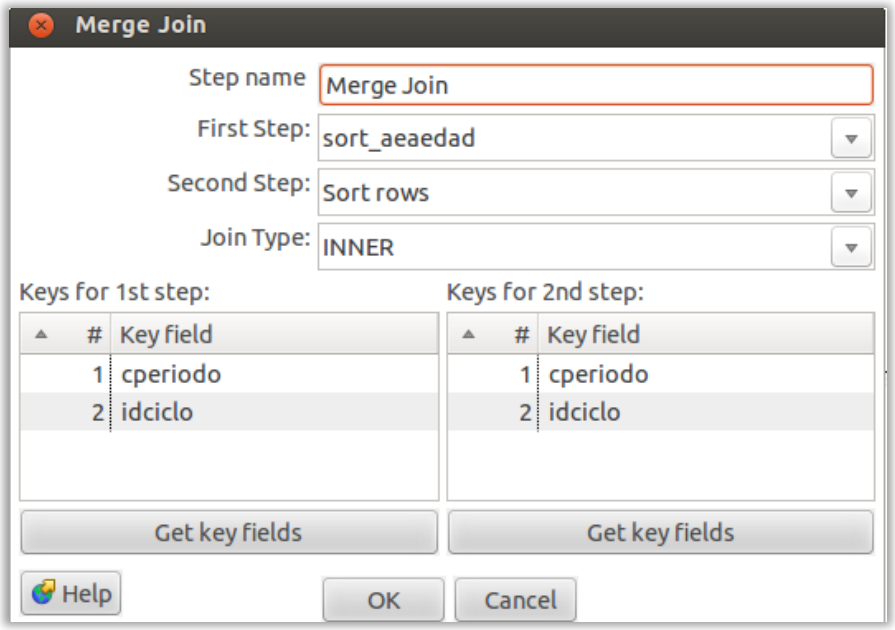
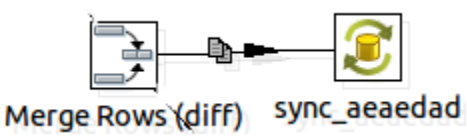


Ilustración 28: Objeto Merge Join

8.2.5.11 Select Values 4

Objeto “Select Values” cuya función es seleccionar los campos cperiodo, idciclo, iduniv, idgnero, idrangoedad, nnummatriculados.

8.2.5.12 Merge Rows (diff) y syn_aeaedad



Estos dos objetos, “Merge Rows (diff)” y “Synchronize after merge” normalmente funcionan combinados.

El primero tiene la función de cruzar dos flujos de datos, devolviendo además un campo flag que indica la estrategia que el objeto “Synchronize after merge” debe seguir posteriormente: insertar, actualizar o eliminar el registro de la tabla destino.

En el objeto “Merge Rows (diff)” es necesario definir cuál es el flujo de datos primario, cuál es el secundario, los campos por los que se cruza, el campo o campos a comparar, y el nombre del campo que contenga el flag.

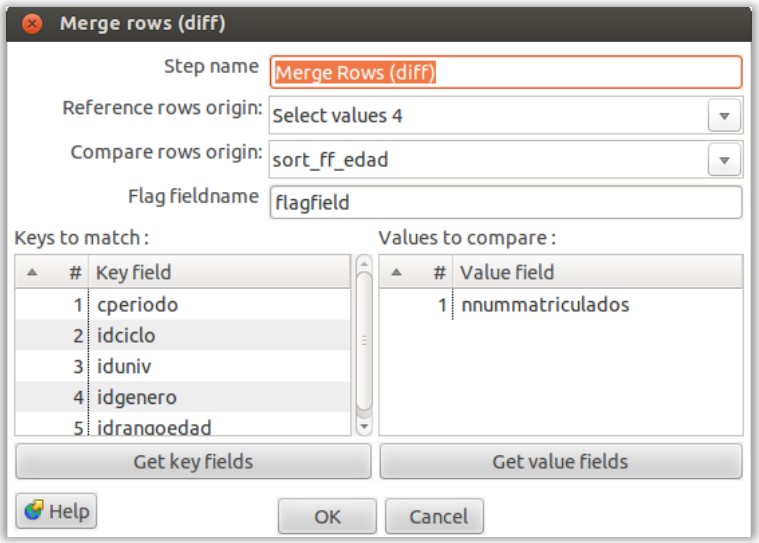


Ilustración 29: Objeto Merge rows(dif)

Por otro lado, el objeto “Synchronize after merge” realiza la operación DML sobre la tabla AEAEDAD. Primero se definen los campos que funcionan como clave en la operación, y después los campos a actualizar o insertar, si procede.

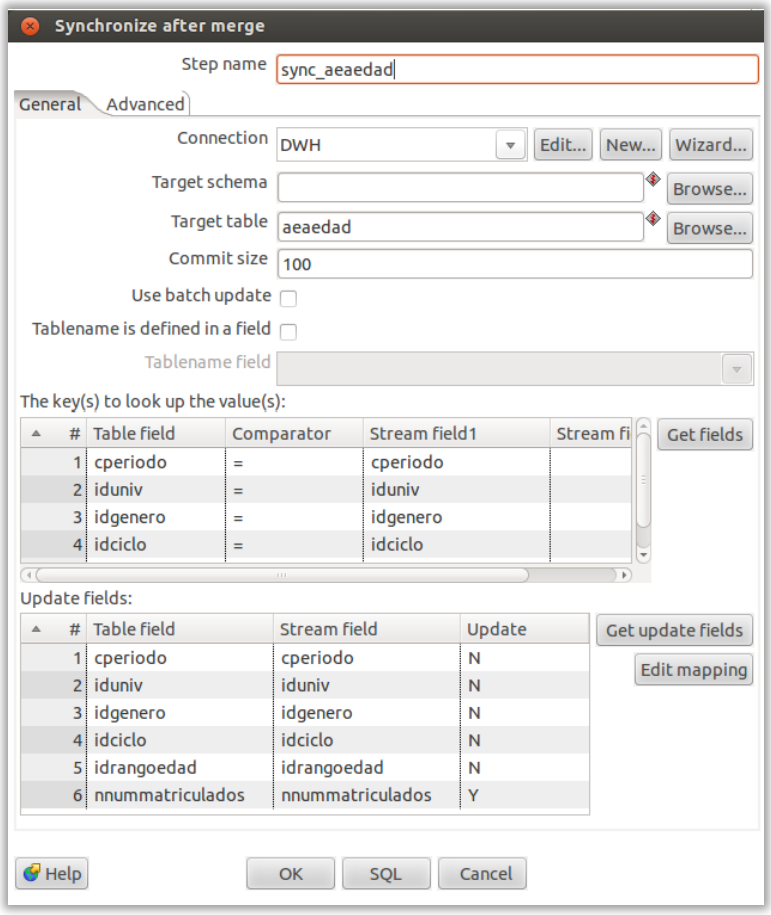


Ilustración 30: Objeto Synchronize after merge - General

Además, se selecciona la estrategia a seguir para cada uno de los valores del flag generado en el objeto "Merge Rows (diff)".

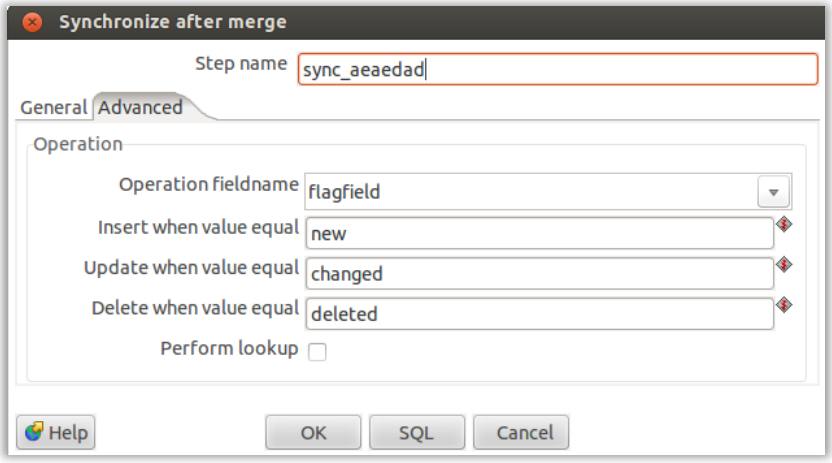


Ilustración 31: Objeto Synchronize after merge - Advanced

8.2.6 t_agregada_rama

Transformación que carga la tabla agregada AEARAMA a partir del fichero "ff_rama.txt". Su funcionamiento es idéntico al de "t_agregada_edad".

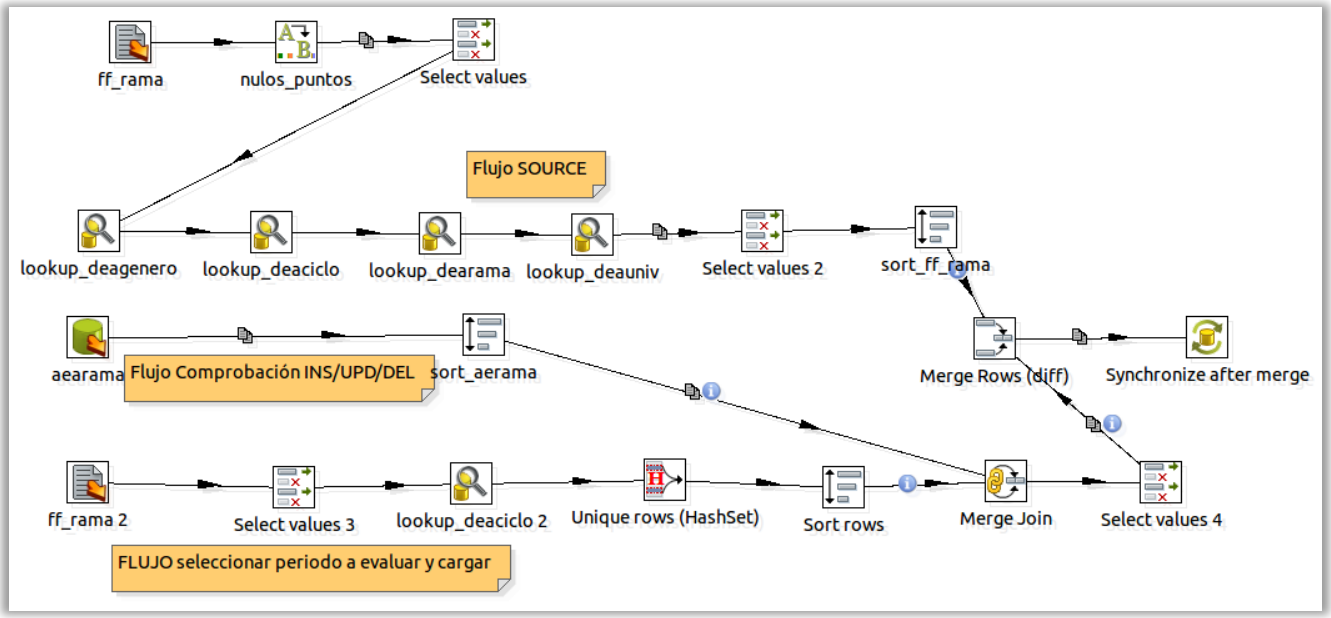


Ilustración 32: Transformación t_agregada_rama

8.2.7 j_Estadisticas_Academicas

Trabajo que secuencialmente ejecuta todas las tareas que intervienen en el proceso ETL. Puede ser lanzado desde la propia herramienta Kettle, o mediante un proceso batch.

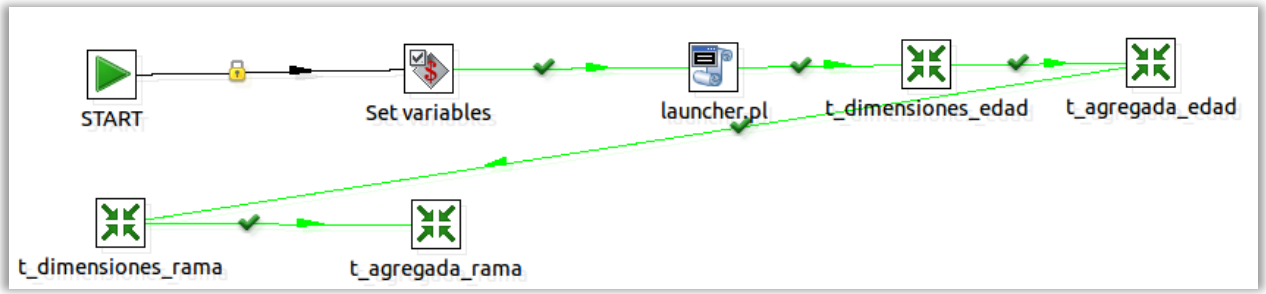


Ilustración 33: Trabajo j_Estadisticas_Academicas

8.2.7.1 Set Variables

Establece el valor de las variables que son comunes en todas las transformaciones.

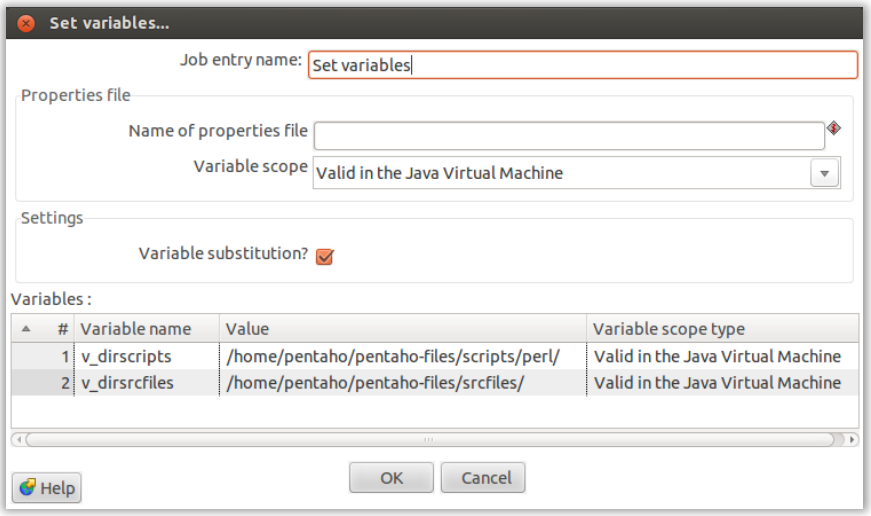


Ilustración 34: Objeto Set variables

8.2.7.2 launcher.pl

Objeto del tipo “Shell” que lanza el proceso desarrollado en Perl, que transforma archivos PC-AXIS en ficheros planos.

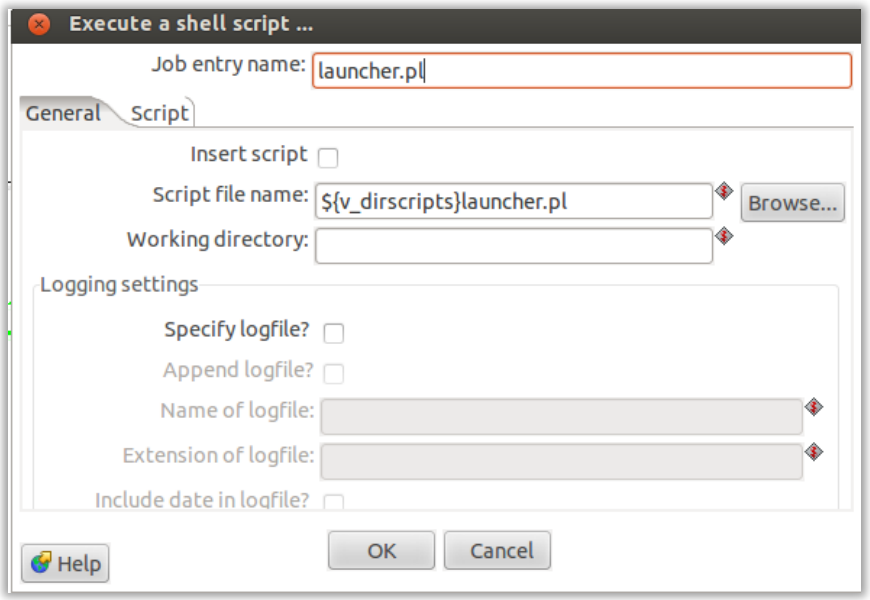


Ilustración 35: Objeto Shell

8.2.7.3 t_dimensiones_edad, t_agregada_edad, t_dimensiones_rama, t_agregada_rama

Objetos "Transformation" que lanzan las transformaciones definidas anteriormente.

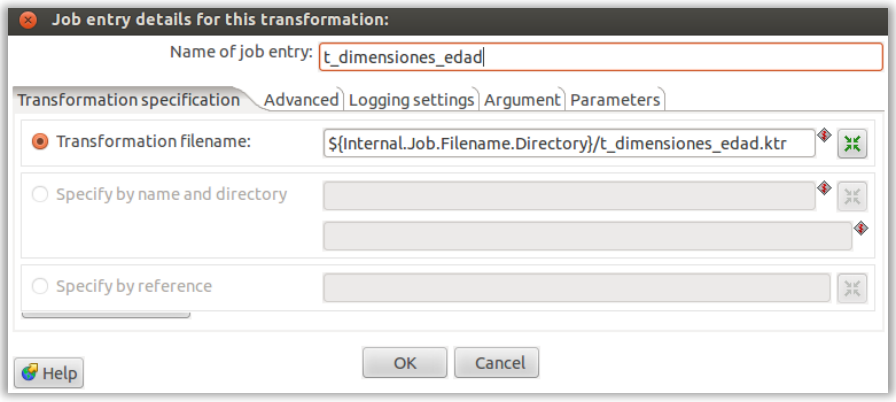


Ilustración 36: Objeto Transformation

8.3 DISEÑO TÉCNICO DE INFORMES

Para el desarrollo de informes del proyecto se utiliza la herramienta Pentaho Report Designer.

Una vez terminados, se publican en el portal web de Pentaho para que sean visibles por todos los usuarios.

El objetivo de los mismos es dar al usuario una visión global de la evolución en el número de alumnos universitarios matriculados, pudiendo desglosar la información por universidades.

Se generan dos informes, uno a nivel nacional y otro con el detalle por universidad:

- 00_Estadistica_Estudiantes_Universitarios
- 01_Estadistica_Universidad

8.3.1 Informe 00_Estadistica_Estudiantes_Universitarios

Informe que contiene el evolutivo del número de alumnos universitarios matriculados y un ranking de las universidades con mayor y menor crecimiento, además de gráficos con información de las principales dimensiones.

Los parámetros de entrada son:

- Periodo (pperiodo)
  Parámetro obligatorio con el periodo de estudio.

- Género (pgenero)
  Parámetro opcional. Los valores posibles son “Hombres”/”Mujeres”

- Ciclo Formativo (pciclo)
  Parámetro opcional que admite múltiples valores. Ciclo formativo en el que está matriculado el alumno (Primer ciclo, segundo ciclo, etc.).

- Universidad (puniv)
  Parámetro opcional que admite múltiples valores. Universidad o universidades de las que se quiere realizar el estudio. Pueden ser tanto públicas como privadas.

Además, se utilizan los parámetros pcheckciclo y pcheckuniv como apoyo para permitir que pciclo y puniv sean opcionales y admitan múltiples valores. Para ello, se marcan como ocultos y se incluye en la opción “Post-Processing Formula” la siguiente fórmula para cada uno de ellos:

=IF(LEN(CSVTEXT([pciclo]))=0; 1; 0)

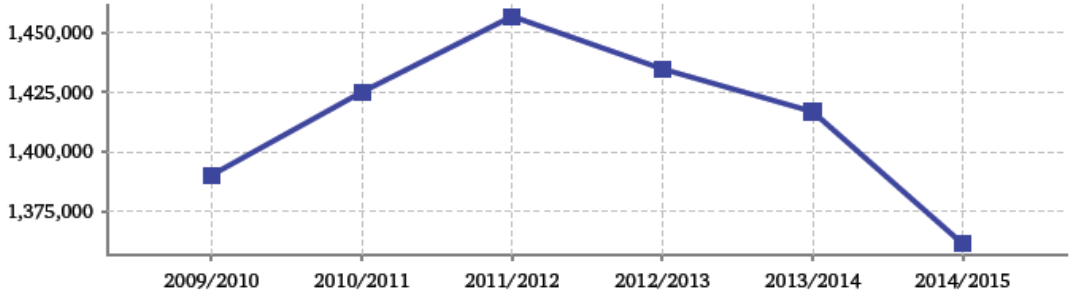
De esta forma, pcheckciclo y pcheckuniv tendrán el valor de 1 si pciclo y puniv no están informados respectivamente.

El informe sigue el siguiente diseño:

Universidad Carlos III de Madrid

GOBIERNO DE ESPAÑA | MINISTERIO DE EDUCACIÓN, CULTURA Y DEPORTE

Evolución número matriculados



Ranking 3 universidades Nº Matriculados

Universidad	Nº Matriculados
Nacional de Educación a Distancia	159,541
Complutense de Madrid	64,492
Sevilla	57,185

Ranking 3 universidades con mayor crecimiento

Universidad	Nº Matr. AÑO ACT.	Nº Matr. AÑO ANT.	Crecimiento
Europea de Canarias	162	101	60.4%
Internacional Valenciana	733	470	56%
IE Universidad	1,409	1,041	35.4%

Ranking 3 universidades con menor crecimiento

Universidad	Nº Matr. AÑO ACT.	Nº Matr. AÑO ANT.	Crecimiento
Oberta de Catalunya	25,403	35,743	(28.9%)
Camilo José Cela	6,343	7,842	(19.1%)
Pablo de Olavide	8,985	10,685	

Ilustración 37: Diseño Informe 00_Estadistica_Estudiantes_Universitarios. Parte 1

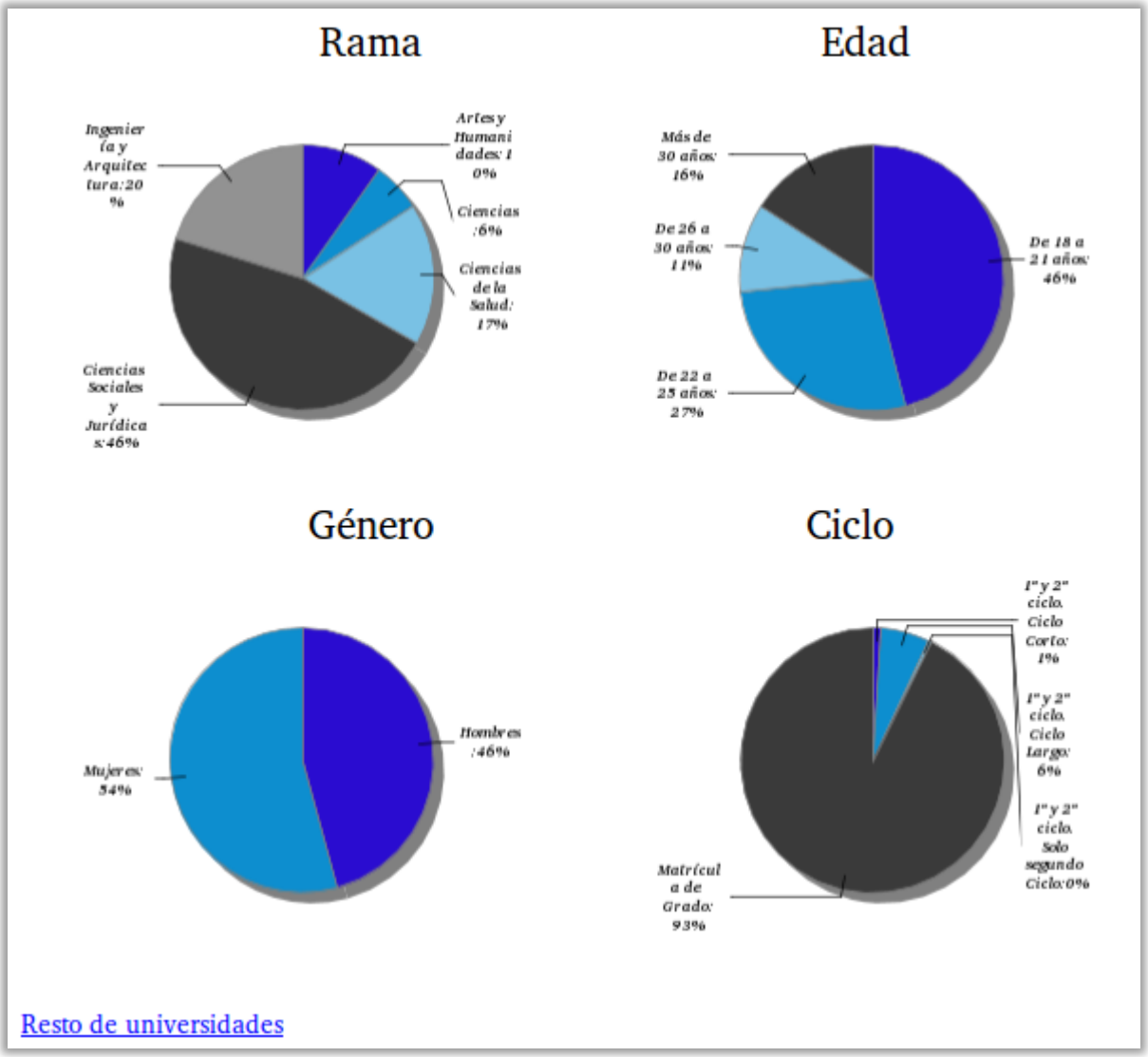


Ilustración 38: Diseño Informe 00_Estadistica_Estudiantes_Universitarios. Parte 2

Pentaho Report Designer solo permite una consulta principal por informe. Si fuera necesario incluir más, éstas tendrían que definirse dentro de subreports a los cuales se les debe pasar también los parámetros de entrada.

En el presente informe se define la consulta utilizada en el primer gráfico de líneas como consulta principal. A los parámetros de entrada pasados a los subreports se les incluye el literal "import" delante de sus nombres, para identificarlos.

Además, los informes están divididos en secciones. Las secciones visibles por defecto son:

- Encabezado de página
- Encabezado de informe

- Detalles
- Pie de informe
- Pie de página

A continuación, se definen los componentes del informe por secciones:

ENCABEZADO DE INFORME:

- Gráfico de Líneas "Evolución Número de Matriculados"
  Parte de una SQL similar a la siguiente:

```
select cperiodo, sum(nnummatriculados) as nummatr
from aeaedad
where
cast(substring(cperiodo, 1, 4) as unsigned) <= cast(substring(${pperiodo}, 1, 4) as unsigned) and
cast(substring(cperiodo, 1, 4) as unsigned) >= cast(substring(${pperiodo}, 1, 4) as unsigned)–5 and
( ${pgenero} is null or idgenero = ${pgenero} ) and
(${pcheckciclo} = 1 or  idciclo     in (${pciclo} )) and
(${pcheckuniv}  = 1 or  iduniv      in (${puniv}  ))
group by cperiodo
```

  El campo cperiodo actúa como "category-colum" y nummatr como "value-columns".

- "Ranking 3 universidades Nº Matriculados"
  Listado con las 3 universidades con mayor número de matriculados, teniendo en cuenta los parámetros de entrada seleccionados. Parte de la siguiente SQL:

```
select
b.cuniv, a.iduniv,
sum(a.nnummatriculados) as nummatr
from
aeaedad a, deauniv b
where
a.iduniv = b.iduniv and
a.cperiodo = ${importperiodo} and
(${importcheckciclo} = 1 or  a.idciclo     in (${importciclo} ))  and
(${importcheckuniv} = 1 or  a.iduniv     in (${importuniv} )) and
(${importgenero} is null or a.idgenero = ${importgenero} )
group by  b.cuniv, a.iduniv
order by nummatr desc
limit 3
```

- "Ranking 3 universidades con mayor crecimiento".
  Listado con las 3 universidades cuyo crecimiento respecto al anterior periodo es el mayor. Se utiliza la siguiente SQL para obtener la información:

```
select
b.cuniv, a.iduniv,
SUM(IF(a.cperiodo= ${importperiodo} ,  a.nnummatriculados, 0) ) as nummatract,
SUM(IF(a.cperiodo= ${importperiodo} , 0,  a.nnummatriculados) ) as nummatrant,
(  SUM(IF(a.cperiodo= ${importperiodo} ,  a.nnummatriculados, 0) ) -
```

```
  SUM(IF(a.cperiodo= ${importperiodo} , 0,  a.nnummatriculados) ) )
 / SUM(IF(a.cperiodo= ${importperiodo} , 0,  a.nnummatriculados) )  as crec
from
aeaedad a, deauniv b
where
a.iduniv = b.iduniv and
( a.cperiodo = ${importperiodo} or substring(a.cperiodo, 1, 4) = cast(substring( ${importperiodo},
1, 4) as unsigned) - 1    ) and
(${importcheckciclo} = 1 or  a.idciclo     in (${importciclo} ))  and
(${importcheckuniv} = 1 or  a.iduniv     in (${importuniv} )) and
(${importgenero} is null or a.idgenero = ${importgenero} )
group by  b.cuniv, a.iduniv
having nummatrant <> 0 and nummatract <> 0
order by crec desc
limit 3
```

- "Ranking 3 universidades con menor crecimiento"
  Listado con las 3 universidades cuyo crecimiento respecto al anterior periodo es el menor. La consulta es exactamente igual que la del componente anterior, salvo que el orden en este caso es ascendente.

PIE DE INFORME

- Gráfico de tarta "Rama".
  Número de matriculados por rama. Se apoya en la siguiente SQL:

```
select b.crama, sum(nnummatriculados) as nummatr
from aearama a, dearama b
where
a.idrama = b.idrama and
a.cperiodo = ${importperiodo} and
(${importcheckciclo} = 1 or  a.idciclo     in (${importciclo} ))  and
(${importcheckuniv} = 1 or  a.iduniv     in (${importuniv} )) and
(${importgenero} is null or a.idgenero = ${importgenero} )
group by 1
```

  El campo crama actúa como "series-by-field" y el campo nummatr como "value-column".

- Gráfico de tarta "Edad".
  Número de matriculados por edad. Utiliza la siguiente SQL:

```
select b.crangoedad, sum(nnummatriculados) as nummatr
from aeaedad a, deaedad b
where
a.idrangoedad = b.idrangoedad and
a.cperiodo = ${importperiodo} and
(${importcheckciclo} = 1 or  a.idciclo     in (${importciclo} ))  and
(${importcheckuniv} = 1 or  a.iduniv     in (${importuniv} )) and
(${importgenero} is null or a.idgenero = ${importgenero} )
group by 1
```

El campo crangoedad actúa como “series-by-field” y el campo nummatr como “value-column”.

- Gráficos de tarta “Género” y “Ciclo”.
  Siguen el mismo funcionamiento que el anterior gráfico, salvo que en este caso se utilizan las dimensiones de género y ciclo respectivamente.

8.3.2 Informe 01_Estadistica_Universidad

Informe que contiene el detalle de la información del número de matriculados por universidad, el evolutivo, y un estudio del crecimiento por dimensión.

Los parámetros de entrada son:

- Universidad (puniv)
  Parámetro obligatorio con la universidad de la que se quiere realizar el estudio.

- Periodo (pperiodo)
  Parámetro obligatorio con el periodo de estudio.

El informe debe seguir el siguiente diseño:

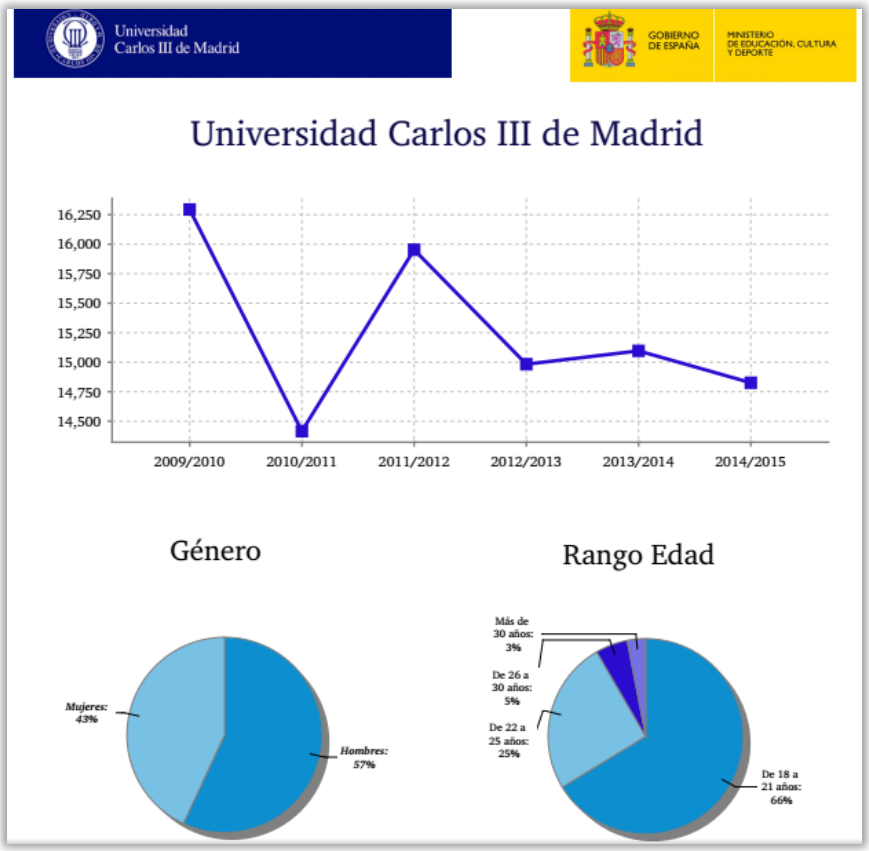
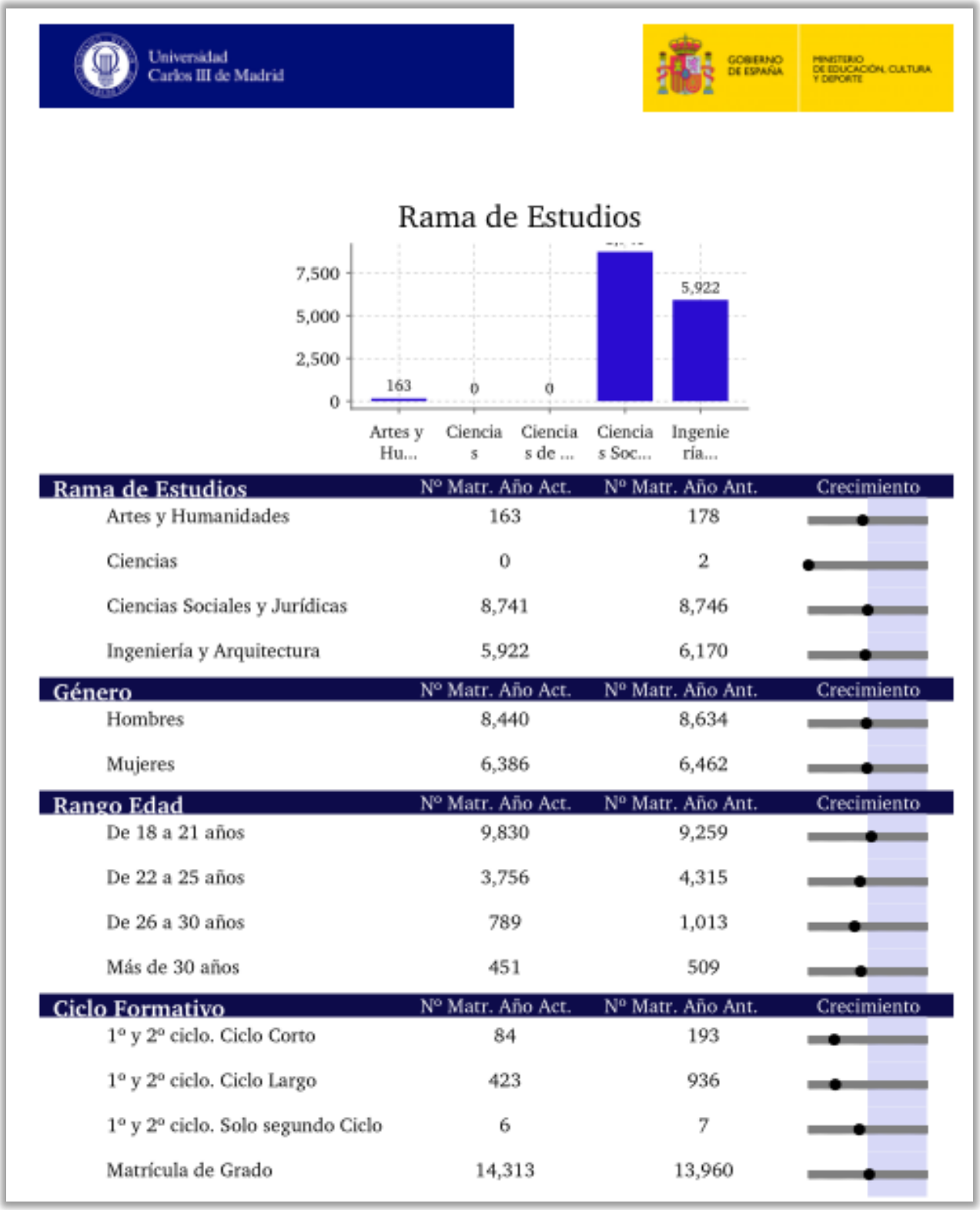


Ilustración 39: Diseño Informe 01_Estadistica_Estadistica_Universidad. Parte 1



Rama de Estudios	Nº Matr. Año Act.	Nº Matr. Año Ant.	Crecimiento
Artes y Humanidades	163	178	
Ciencias	0	2	
Ciencias Sociales y Jurídicas	8,741	8,746	
Ingeniería y Arquitectura	5,922	6,170	
Género	**Nº Matr. Año Act.**	**Nº Matr. Año Ant.**	**Crecimiento**
Hombres	8,440	8,634	
Mujeres	6,386	6,462	
Rango Edad	**Nº Matr. Año Act.**	**Nº Matr. Año Ant.**	**Crecimiento**
De 18 a 21 años	9,830	9,259	
De 22 a 25 años	3,756	4,315	
De 26 a 30 años	789	1,013	
Más de 30 años	451	509	
Ciclo Formativo	**Nº Matr. Año Act.**	**Nº Matr. Año Ant.**	**Crecimiento**
1º y 2º ciclo. Ciclo Corto	84	193	
1º y 2º ciclo. Ciclo Largo	423	936	
1º y 2º ciclo. Solo segundo Ciclo	6	7	
Matrícula de Grado	14,313	13,960	

Ilustración 40: Diseño Informe 01_Estadistica_Estadistica_Universidad. Parte 2

La consulta principal del informe es la que calcula el crecimiento por dimensiones. El resto de objetos están incrustados en subreports.

ENCABEZADO DE PÁGINA:

- Título del informe.
  Se define el título del informe a partir del parámetro de entrada seleccionado. Para ello se utiliza una función que se apoye en el resultado de una consulta sobre base de datos:

  =SINGLEVALUEQUERY("display_univ")

  Esta función devuelve el primer registro devuelto en la consulta seleccionada, cuya SQL es la siguiente:

  Select cuniv from deauniv where iduniv = ${puniv}

ENCABEZADO DE INFORME:

- Gráfico de líneas con la evolución del número de matriculados de la universidad seleccionada. Parte de la siguiente SQL:

  ```
  select b.cuniv, a.cperiodo, sum(nnummatriculados) as nummatr
  from aeaedad a, deauniv b
  where
  a.iduniv = b.iduniv and
  a.iduniv = ${importuniv} and
  cast(substring(a.cperiodo, 1, 4) as unsigned) <= cast(substring(${importperiodo}, 1, 4) as
  unsigned) and
  cast(substring(a.cperiodo, 1, 4) as unsigned) >= cast(substring(${importperiodo}, 1, 4) as
  unsigned) - 5
  group by 1, 2
  ```

  El campo cperiodo actúa como "category-column" y nummatr como "value-columns".

- Gráfico de tarta de las dimensiones género y rango de edad.
  Estos gráficos son muy similares a los definidos en el anterior informe. Las SQL de las que parten son:

  ```
  select b.cgenero, sum(nnummatriculados) as nummatr
  from aeaedad a, deagenero b
  where
  a.idgenero = b.idgenero and
  a.iduniv = ${importuniv} and
  a.cperiodo = ${importperiodo}
  group by 1

  select b.crangoedad, sum(nnummatriculados) as nummatr
  from aeaedad a, deaedad b
  where
  a.idrangoedad = b.idrangoedad and
  a.iduniv = ${importuniv} and
  a.cperiodo = ${importperiodo}
  ```

```
group by 1
```

- Gráfico de Barras de la dimensión ciclo formativo.
  Parte de una consulta similar a la siguiente:

```
select b.cciclo, sum(nnummatriculados) as nummatr
from aeaedad a, deaciclo b
where
a.idciclo = b.idciclo and
a.iduniv = ${importuniv} and
a.cperiodo = ${importperiodo}
group by 1
```

  El campo cciclo actúa como “category-column”, y nummatr como “value-columns”.

DETALLES:

- Detalle del crecimiento por dimensión.
  Se desglosan los valores de cada dimensión, incluyendo las medidas del número de matriculados en el periodo seleccionado, el número de matriculados en el periodo anterior, y el crecimiento, mostrándolo mediante un objeto “survey-scale”.
  Los objetos “survey-scale” son gráficos muy pequeños que representan un valor dentro de una escala. En este caso, la escala es un porcentaje que puede ir desde -100% hasta el 100%.
  El listado parte de una SQL similar a la siguiente:

```
select
'Género' as dimension,
b.cgenero cvalor,
SUM(IF(a.cperiodo= ${pperiodo} ,  a.nnummatriculados, 0) ) as act,
SUM(IF(a.cperiodo= ${pperiodo} , 0,  a.nnummatriculados) ) as ant,
(SUM(IF(a.cperiodo= ${pperiodo} ,  a.nnummatriculados, 0) ) - SUM(IF(a.cperiodo= ${pperiodo} ,
0,  a.nnummatriculados) ) ) / SUM(IF(a.cperiodo= ${pperiodo} , 0,  a.nnummatriculados) ) *100 as
crec
from
aeaedad a, deagenero b
where
a.idgenero = b.idgenero and
( a.cperiodo = ${pperiodo} or substring(a.cperiodo, 1, 4) = cast(substring( ${pperiodo}, 1, 4) as
unsigned) - 1    ) and
a.iduniv = ${puniv}
group by  dimension, b.cgenero
having ant <> 0
union
… <el resto de la consulta sigue el mismo patrón, cambiando de dimensión en cada caso>
```

8.3.3 Publicar informes en el portal web

Una vez desarrollados y guardados los informes, éstos se publican en el portal web de Pentaho, para que sean visibles por todos los usuarios de la compañía, y además, puedan ser vinculables entre ellos.

Se puede publicar un informe, dentro de la herramienta Pentaho Report Designer, en el menú File / Publish.

El usuario y contraseña del portal BI de Pentaho del proyecto, es el que viene por defecto en la instalación de la herramienta, “admin”/”password”.

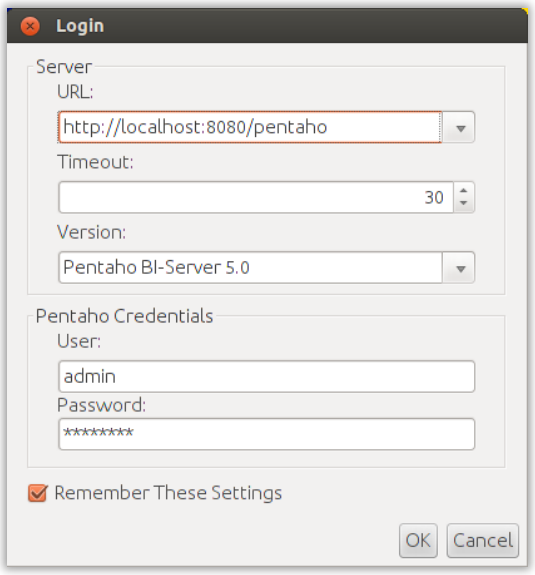


Ilustración 41: Publicar informes servidor Pentaho. Login

Posteriormente, se elige el nombre y ubicación donde se desea publicar.

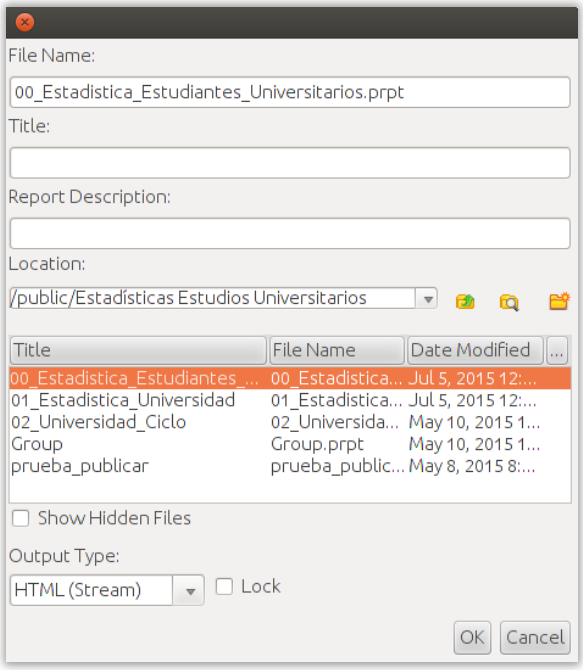


Ilustración 42: Publicar informes servidor Pentaho. Ubicación

Se puede acceder al portal web de Pentaho, dentro de la máquina virtual, a través del enlace:

http://localhost:8080/pentaho

Una vez dentro, el portal web de Pentaho tiene el siguiente aspecto:

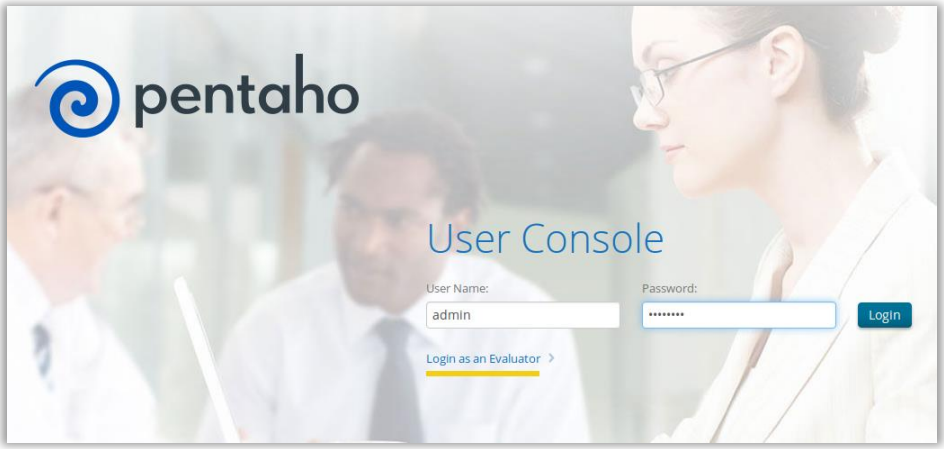


Ilustración 43: Portal web Pentaho. Login

Ilustración 44:Portal web Pentaho. Home

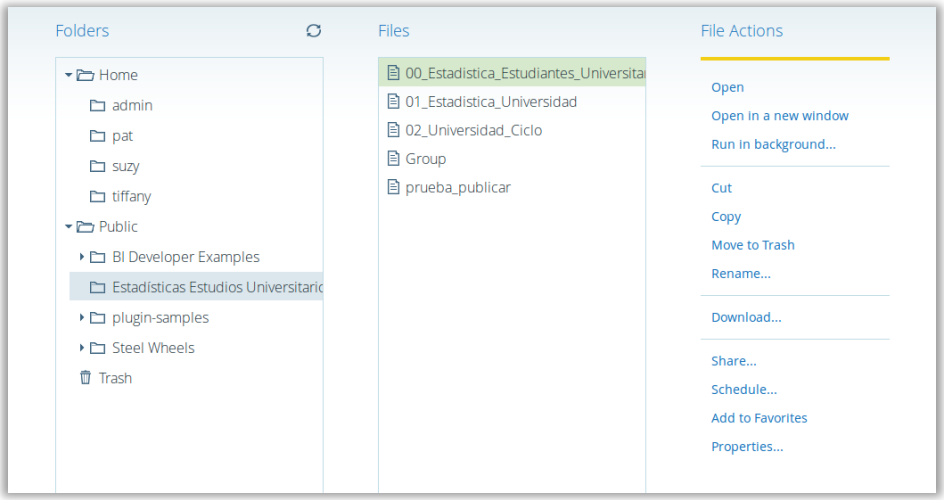


Ilustración 45: Portal web Pentaho. Explorar Archivos

8.3.4 Vincular informes

Después de publicar los informes en el portal web, es posible vincularlos entre ellos.

En concreto, se incluyen vínculos en el informe "00_Estadistica_Estudiantes_Universitarios", para que al pinchar en alguna universidad, se abra el informe "01_Estadistica_Universidad" pasándose como parámetros el periodo y la universidad seleccionada.

Para ello, dentro de la herramienta Pentaho Report Designer, se selecciona el objeto al que se desea añadir el vínculo, y se pincha en el menú contextual la opción "Hyperlink":

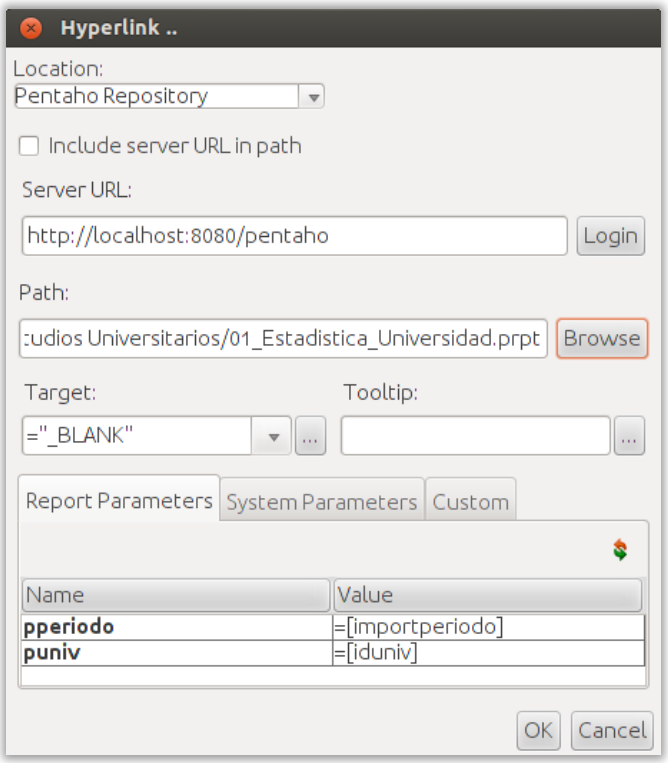


Ilustración 46: Hipervínculos a informes

9 PLAN DE PRUEBAS

A continuación se detallan los casos de pruebas que verifican que el sistema satisface los requisitos especificados.

9.1 PLAN DE PRUEBAS DE LA ETL

9.1.1 Pruebas t_dimensiones_edad

Nº	Descripción Prueba	Resultado Esperado	Resultado obtenido	Querys
1	La ejecución unitaria de la transformación t_dimensiones_edad finaliza sin errores.	0 errores.	El esperado	
2	El número de registros cargados en la tabla DEACICLO es el correcto.	5 registros.	El esperado	select count(*) from DEACICLO
3	El número de registros cargados en la tabla DEAGENERO es el correcto.	2 registros.	El esperado	select count(*) from DEAGENERO
4	El número de registros cargados en la tabla DEAEDAD es el correcto.	4 registros.	El esperado	select count(*) from DEAEDAD
5	El número de registros cargados en la tabla DEAUNIV es el correcto.	83 registros.	El esperado	select count(*) from DEAUNIV
6	El número de registros cargados en la tabla DEATIPUNIV es el correcto.	2 registros.	El esperado	select count(*) from DEATIPUNIV
7	Una ejecución posterior no altera los valores, ya cargados, de las tablas.	0 actualizaciones.	El esperado	
8	Se eliminan algunos registros de prueba, de cada una de las tablas. El proceso debe volver a cargarlos, aunque con diferentes secuenciales, ya que éstos se calculan con el máximo identificador + 1.	Inserción de los registros eliminados previamente.	El esperado	
9	El proceso no borra registros. Para probar, se insertan registros ficticios, que no existan en los ficheros de origen. El proceso no debe borrarlos.	0 borrados.	El esperado	
10	El contenido de las tablas es correcto, comprobando que no se trunca ningún campo, y que no hay caracteres inválidos.		El esperado	

Tabla 34: Pruebas t_dimensiones_edad

9.1.2 Pruebas t_dimensiones_rama

Nº	Descripción Prueba	Resultado Esperado	Resultado obtenido	Querys
1	La ejecución unitaria de la transformación t_dimensiones_rama finaliza sin errores.	0 errores.	El esperado	
2	El número de registros cargados en la tabla DEACICLO es el correcto.	5 registros.	El esperado	select count(*) from DEACICLO
3	El número de registros cargados en la tabla DEAGENERO es el correcto.	2 registros.	El esperado	select count(*) from DEAGENERO
4	El número de registros cargados en la tabla DEARAMA es el correcto.	7 registros.	El esperado	select count(*) from DEARAMA
5	El número de registros cargados en la tabla DEAUNIV es el correcto.	83 registros.	El esperado	select count(*) from DEAUNIV
6	El número de registros cargados en la tabla DEATIPUNIV es el correcto.	2 registros.	El esperado	select count(*) from DEATIPUNIV
7	Una ejecución posterior no altera los valores, ya cargados, de las tablas.	0 actualizaciones.	El esperado	
8	Se eliminan algunos registros de prueba, de cada una de las tablas. El proceso debe volver a cargarlos, aunque con diferentes secuenciales, ya que éstos se calculan con el máximo identificador + 1.	Inserción de los registros eliminados previamente.	El esperado	
9	El proceso no borra registros. Para probar, se insertan registros ficticios, que no existan en los ficheros de origen. El proceso no debe borrarlos.	0 borrados.	El esperado	
10	El contenido de las tablas es correcto, comprobando que no se trunca ningún campo, y que no hay caracteres inválidos.		El esperado	

Tabla 35: Pruebas t_dimensiones_rama

9.1.3 Pruebas t_agregada_edad

Nº	Descripción Prueba	Resultado Esperado	Resultado obtenido	Querys
1	La ejecución unitaria de la transformación t_agregada_edad finaliza sin errores.	0 errores.	El esperado	
2	El número de registros cargados en la tabla AEAEDAD es el correcto. Los periodos de carga utilizados en las pruebas van desde 2008/2009 hasta 2014/2015.	14748 registros.	El esperado	select count(*) from AEAEDAD

3	El número de alumnos matriculados por periodo es el correcto. Para comprobarlo, contrastar el dato con el publicado en la web del Ministerio de Educación.	1361340,00 2014/2015 1416827,00 2013/2014 1434729,00 2012/2013 1456783,00 2011/2012 1425018,00 2010/2011 1390234,00 2009/2010 1379726,00 2008/2009	El esperado	select sum(nnummatriculados), cperiodo from aeaedad group by cperiodo
4	El número de alumnos matriculados por género y tipo de universidad, para el periodo 2014/2015 es el correcto. Contrastar con datos de la web del Ministerio de Educación.	73584,00 Privada Hombres 91192,00 Privada Mujeres 548413,00 Pública Hombres 648151,00 Pública Mujeres	El esperado	select sum(nnummatriculados), d.ctipuniv, b.cgenero from aeaedad a, deagenero b, deauniv c, deatipuniv d where a.idgenero = b.idgenero and a.iduniv = c.iduniv and c.idtipuniv = d.idtipuniv and a.cperiodo = '2014/2015' group by 2, 3
5	El número de alumnos matriculados por edad y tipo de universidad, para el periodo 2014/2015 es el correcto. Contrastar con datos de la web del Ministerio de Educación.	64064,00 Privada De 18 a 21 años 42224,00 Privada De 22 a 25 años 21220,00 Privada De 26 a 30 años 37268,00 Privada Más de 30 años 560017,00 Pública De 18 a 21 años 330253,00 Pública De 22 a 25 años 125832,00 Pública De 26 a 30 años 180462,00 Pública Más de 30 años	El esperado	select sum(nnummatriculados), d.ctipuniv, b.crangoedad from aeaedad a, deaedad b, deauniv c, deatipuniv d where a.idrangoedad = b.idrangoedad and a.iduniv = c.iduniv and c.idtipuniv = d.idtipuniv and a.cperiodo = '2014/2015' group by 2, 3

6	El número de alumnos matriculados por ciclo y tipo de universidad, para el periodo 2014/2015 es el correcto. Contrastar con datos de la web del Ministerio de Educación.	315,00 Privada 1º y 2º ciclo, Ciclo Corto 3589,00 Privada 1º y 2º ciclo, Ciclo Largo 167,00 Privada 1º y 2º ciclo, Solo segundo Ciclo 160705,00 Privada Matrícula de Grado 15267,00 Pública 1º y 2º ciclo, Ciclo Corto 76033,00 Pública 1º y 2º ciclo, Ciclo Largo 5443,00 Pública 1º y 2º ciclo, Solo segundo Ciclo 1099821,00 Pública Matrícula de Grado		select sum(nnummatriculados), d.ctipuniv, b.cciclo from aeaedad a, deaciclo b, deauniv c, deatipuniv d where a.idciclo = b.idciclo and a.iduniv = c.iduniv and c.idtipuniv = d.idtipuniv and a.cperiodo = '2014/2015' group by 2, 3
7	El número de alumnos matriculados por universidad y tipo de universidad, para el periodo 2014/2015 es el correcto. Contrastar con datos de la web del Ministerio de Educación.	159541 Pública Nacional de Educación a Distancia 64492 Pública Complutense de Madrid 57185 Pública Sevilla 49888 Pública Granada 45101 Pública Barcelona 41073 Pública Valiéncia (Estudi General) 39099 Pública País Vasco/Euskal Herriko Unibertsitatea 35737 Pública Rey Juan Carlos 34065 Pública Málaga 33453 Pública Autónoma de Barcelona 32849 Pública Politécnica de Madrid 28754 Pública Zaragoza 28492 Pública Murcia		select sum(nnummatriculados), d.ctipuniv, c.cuniv from aeaedad a, deauniv c, deatipuniv d where a.iduniv = c.iduniv and c.idtipuniv = d.idtipuniv and a.cperiodo = '2014/2015' group by 2, 3
8	Una ejecución posterior no altera los valores, ya cargados, si no hay cambios en el fichero de origen.	0 actualizados.	El esperado	
9	El proceso inserta/elimina/actualiza correctamente, si cambia algún valor del fichero origen, sólo para el periodo y ciclo del fichero de entrada. Para probar, se eliminan, insertan y actualizan registros de la tabla AEAEDAD, con valores ficticios.	El proceso debe dejarlos en su estado inicial.	El esperado	

Tabla 36: Pruebas t_agregada_edad

9.1.4 Pruebas t_agregada_rama

Nº	Descripción Prueba	Resultado Esperado	Resultado obtenido	Querys
1	La ejecución unitaria de la transformación t_agregada_rama finaliza sin errores.	0 errores.	El esperado	
2	El número de registros cargados en la tabla AEARAMA es el correcto. Los periodos de carga utilizados en las pruebas van desde 2008/2009 hasta 2014/2015.	16774 registros.	El esperado	select count(*) from AEARAMA
3	El número de alumnos matriculados por periodo es el correcto. Para comprobarlo, contrastar el dato con el publicado en la web del Ministerio de Educación, y con aeaedad.	1361340,00 2014/2015 1416827,00 2013/2014 1434729,00 2012/2013 1456783,00 2011/2012 1425018,00 2010/2011 1390234,00 2009/2010 1379726,00 2008/2009	El esperado	select sum(nnummatriculados), cperiodo from aearama group by cperiodo
4	El número de alumnos matriculados por género y tipo de universidad, para el periodo 2014/2015 es el correcto. Contrastar con datos de la web del Ministerio de Educación, y con aeaedad.	73584,00 Privada Hombres 91192,00 Privada Mujeres 548413,00 Pública Hombres 648151,00 Pública Mujeres	El esperado	select sum(nnummatriculados), d.ctipuniv, b.cgenero from aearama a, deagenero b, deauniv c, deatipuniv d where a.idgenero = b.idgenero and a.iduniv = c.iduniv and c.idtipuniv = d.idtipuniv and a.cperiodo = '2014/2015' group by 2, 3

5	El número de alumnos matriculados por rama y tipo de universidad, para el periodo 2014/2015 es el correcto. Contrastar con datos de la web del Ministerio de Educación.	6619,00 Privada Artes y Humanidades 2019,00 Privada Ciencias 46902,00 Privada Ciencias de la Salud 89913,00 Privada Ciencias Sociales y Jurídicas 19323,00 Privada Ingeniería y Arquitectura 128621,00 Pública Artes y Humanidades 79313,00 Pública Ciencias 189959,00 Pública Ciencias de la Salud 543018,00 Pública Ciencias Sociales y Jurídicas 255653,00 Pública Ingeniería y Arquitectura	El esperado	select sum(nnummatriculados), d.ctipuniv, b.crama from aearama a, dearama b, deauniv c, deatipuniv d where a.idrama = b.idrama and a.iduniv = c.iduniv and c.idtipuniv = d.idtipuniv and a.cperiodo = '2014/2015' group by 2, 3
6	El número de alumnos matriculados por ciclo y tipo de universidad, para el periodo 2014/2015 es el correcto. Contrastar con datos de la web del Ministerio de Educación, y con aeaedad.	315,00 Privada 1º y 2º ciclo, Ciclo Corto 3589,00 Privada 1º y 2º ciclo, Ciclo Largo 167,00 Privada 1º y 2º ciclo, Solo segundo Ciclo 160705,00 Privada Matrícula de Grado 15267,00 Pública 1º y 2º ciclo, Ciclo Corto 76033,00 Pública 1º y 2º ciclo, Ciclo Largo 5443,00 Pública 1º y 2º ciclo, Solo segundo Ciclo 1099821,00 Pública Matrícula de Grado		select sum(nnummatriculados), d.ctipuniv, b.cciclo from aearama a, deaciclo b, deauniv c, deatipuniv d where a.idciclo = b.idciclo and a.iduniv = c.iduniv and c.idtipuniv = d.idtipuniv and a.cperiodo = '2014/2015' group by 2, 3
7	El número de alumnos matriculados por universidad y tipo de universidad, para el periodo 2014/2015 es el correcto. Contrastar con datos de la web del Ministerio de Educación, y con aeaedad.	159541 Pública Nacional de Educación a Distancia 64492 Pública Complutense de Madrid 57185 Pública Sevilla 49888 Pública Granada 45101 Pública Barcelona 41073 Pública Valiéncia (Estudi General) 39099 Pública País Vasco/Euskal Herriko Unibertsitatea 35737 Pública Rey Juan Carlos 34065 Pública Málaga 33453 Pública Autónoma de Barcelona 32849 Pública Politécnica de Madrid 28754 Pública Zaragoza 28492 Pública Murcia		select sum(nnummatriculados), d.ctipuniv, c.cuniv from aearama a, deauniv c, deatipuniv d where a.iduniv = c.iduniv and c.idtipuniv = d.idtipuniv and a.cperiodo = '2014/2015' group by 2, 3

8	Una ejecución posterior no altera los valores, ya cargados, si no hay cambios en el fichero de origen.		El esperado	
9	El proceso inserta/elimina/actualiza correctamente, si cambia algún valor del fichero origen, sólo para el periodo y ciclo del fichero de entrada. Para probar, se eliminan, insertan y actualizan registros de la tabla AEARAMA, con valores ficticios.	El proceso debe dejarlos en su estado inicial.	El esperado	

Tabla 37: Pruebas t_agregada_rama

9.1.5 Pruebas j_Estadisticas_Academicas

Nº	Descripción Prueba	Resultado Esperado	Resultado obtenido	Querys
1	La ejecución integrada del trabajo j_Estadisticas_Academicas finaliza sin errores.	0 errores.	El esperado	
2	El job establece las variables de proceso correctamente.	v_dirscripts = /home/pentaho/pentaho-files/scripts/perl/ Directorio con los ficheros de entrada v_dirsrcfiles = /home/pentaho/pentaho-files/srcfiles/	El esperado	
3	Comprobar la correcta ejecución del script de Perl, generando los ficheros de salida ff_edad.txt y ff_rama.txt.		El esperado	
4	El proceso ejecuta todas las transformaciones correctamente de forma secuencial.		El esperado	

Tabla 38: Pruebas j_Estadisticas_Academicas

9.2 PLAN DE PRUEBAS DE INFORMES

9.2.1 Pruebas 00_Estadistica_Estudiantes_Universitarios

Nº	Descripción Prueba	Resultado Esperado	Resultado obtenido	Querys
1	Los datos mostrados en el informe son correctos. Cuadran con los valores de las respectivas tablas en BBDD.		El esperado	
2	El informe filtra correctamente por los parámetros de entrada seleccionados: - Periodo - Género - Ciclo Formativo - Universidad		El esperado	
3	Comprobar que calcula correctamente las 3 universidades con mayor número de matriculados. Para las pruebas se selecciona el periodo 2014/2015.	159541 Nacional de Educación a Distancia 64492 Complutense de Madrid 57185 Sevilla	El esperado	b.cuniv, a.iduniv, sum(a.nnummatriculados) as nummatr from aeaedad a, deauniv b where a.iduniv = b.iduniv and a.cperiodo = '2014/2015' group by b.cuniv, a.iduniv order by nummatr desc limit 3

4	Comprobar que calcula correctamente el ranking de las 3 universidades con mayor crecimiento en el periodo. Para ello, se selecciona el periodo 2014/2015.	Europea de Canarias 60.4% Internacional Valenciana 56% IE Universidad 35.4%	El esperado	select b.cuniv, a.iduniv, SUM(IF(a.cperiodo= '2014/2015' , a.nnummatriculados, 0)) as nummatract, SUM(IF(a.cperiodo= '2014/2015' , 0, a.nnummatriculados)) as nummatrant, (SUM(IF(a.cperiodo= '2014/2015' , a.nnummatriculados, 0)) - SUM(IF(a.cperiodo= '2014/2015' , 0, a.nnummatriculados))) / SUM(IF(a.cperiodo= '2014/2015' , 0, a.nnummatriculados)) as crec from aeaedad a, deauniv b where a.iduniv = b.iduniv and a.cperiodo in ('2014/2015', '2013/2014') group by b.cuniv, a.iduniv having nummatrant <> 0 and nummatract <> 0 order by crec desc limit 3
5	Comprobar que calcula correctamente el ranking de las 3 universidades con menor crecimiento en el periodo. Para ello, se selecciona el periodo 2014/2015.	Oberta de Catalunya -28.9% Camilo José Cela -19.1% Pablo de Olavide -15.9%	El esperado	select b.cuniv, a.iduniv, SUM(IF(a.cperiodo= '2014/2015' , a.nnummatriculados, 0)) as nummatract, SUM(IF(a.cperiodo= '2014/2015' , 0, a.nnummatriculados)) as nummatrant, (SUM(IF(a.cperiodo= '2014/2015' , a.nnummatriculados, 0)) - SUM(IF(a.cperiodo= '2014/2015' , 0, a.nnummatriculados))) / SUM(IF(a.cperiodo= '2014/2015' , 0, a.nnummatriculados)) as crec from aeaedad a, deauniv b where a.iduniv = b.iduniv and a.cperiodo in ('2014/2015', '2013/2014') group by b.cuniv, a.iduniv having nummatrant <> 0 and nummatract <> 0 order by crec asc limit 3

6	Los gráficos de tarta por dimensiones muestran los porcentajes correctamente.		El esperado	
7	Los hipervínculos al informe 01 con el detalle de la universidad son correctos, pasando los parámetros de universidad y periodo.		El esperado	
8	Comprobar los diferentes formatos de salida: HTML, PDF, csv.		El esperado	

Tabla 39: Pruebas 00_Estadistica_Estudiantes_Universitarios

9.2.2 Pruebas 01_Estadistica_Universidad

Nº	Descripción Prueba	Resultado Esperado	Resultado obtenido	Querys
1	Los datos mostrados en el informe son correctos. Cuadran con los valores de las respectivas tablas en BBDD.		El esperado	
2	El informe filtra correctamente por los parámetros de entrada seleccionados: - Periodo - Universidad		El esperado	
3	Los gráficos de tarta por dimensión son correctos.		El esperado	

4	Comprobar que el detalle de crecimiento por dimensiones es correcto. Se prueba para el periodo 2014/2015, Universidad Carlos III de Madrid.	Género Hombres -2,25% Mujeres -1,18% Rango Edad De 18 a 21 años 6,17% De 22 a 25 años -12,95% De 26 a 30 años -22,11% Más de 30 años -11,39% Ciclo Formativo 1º y 2º ciclo. Ciclo Corto -56,48% 1º y 2º ciclo. Ciclo Largo -54,81% 1º y 2º ciclo. Solo segundo Ciclo -14,29% Matrícula de Grado 2,53%	El esperado	select 'Género' as dimension, b.cgenero cvalor, (SUM(IF(a.cperiodo= '2014/2015' , a.nnummatriculados, 0)) - SUM(IF(a.cperiodo= '2014/2015' , 0, a.nnummatriculados))) / SUM(IF(a.cperiodo= '2014/2015' , 0, a.nnummatriculados)) *100 as crec from aeaedad a, deagenero b where a.idgenero = b.idgenero and (a.cperiodo in ('2014/2015', '2103/2014') and a.iduniv = 'Carlos III de Madrid' group by dimension, b.cgenero having ant <> 0 union select 'Rango Edad' as dimension, b.crangoedad cvalor, (SUM(IF(a.cperiodo= '2014/2015' , a.nnummatriculados, 0)) - SUM(IF(a.cperiodo= '2014/2015' , 0, a.nnummatriculados))) / SUM(IF(a.cperiodo= '2014/2015' , 0, a.nnummatriculados)) *100 as crec from aeaedad a, deaedad b where a.idrangoedad = b.idrangoedad and (a.cperiodo in ('2014/2015', '2103/2014') and a.iduniv = 'Carlos III de Madrid' group by dimension, b.crangoedad having ant <> 0 union select 'Ciclo Formativo' as dimension, b.cciclo cvalor, (SUM(IF(a.cperiodo= '2014/2015' , a.nnummatriculados, 0)) - SUM(IF(a.cperiodo= '2014/2015' , 0, a.nnummatriculados))) / SUM(IF(a.cperiodo= '2014/2015' , 0, a.nnummatriculados)) *100 as crec from aeaedad a, deaciclo b

				where a.idciclo = b.idciclo and (a.cperiodo in ('2014/2015', '2103/2014') and a.iduniv = 'Carlos III de Madrid' group by dimension, b.cciclo having ant <> 0
5	Comprobar los diferentes formatos de salida: HTML, PDF, csv.		El esperado	

Tabla 40: Pruebas 01_Estadistica_Universidad

10 CONCLUSIÓN

El principal objetivo del proyecto era estudiar la viabilidad del desarrollo de plataformas Business Intelligence orientadas a PYMES, teniendo en cuenta el bajo presupuesto del que disponen.

El presupuesto del proyecto ha ascendido a la cifra de 6.718 €. Aunque el coste es mínimo, no se puede considerar una cifra asequible para una PYME.

Pero hay que valorar varios factores:

- La curva de aprendizaje de las herramientas de Pentaho es exponencial. El mayor grueso de la dedicación en el proyecto se ha invertido en aprendizaje y estudio de la herramienta. Esas horas no serán necesarias en proyectos futuros si no se modifican los recursos, o se contratan recursos experimentados.
- El objetivo de todo proyecto no es solamente obtener beneficios. Hay proyectos que buscan abrirse a nuevos clientes, o estudiar las posibilidades de un nuevo mercado, como es el caso.
- El proyecto no necesita de ninguna licencia, por lo que todo el presupuesto va dirigido al esfuerzo de los recursos.

Analizando estas consideraciones, se podría decir que el desarrollo de proyectos Business Intelligence orientados a PYMES es una buena oportunidad de negocio, ya que es un mercado muy poco explotado hasta ahora.

Las herramientas de Pentaho han demostrado ser potentes, fiables, bien integradas, y con una gran facilidad de instalación y puesta en marcha. El único inconveniente que se le encuentra es un diseño de informes un poco pobre, que aunque se pueda suplir con la incorporación de extensiones ofrecidas por la comunidad, sería más práctico que las mismas ya estuvieran incorporadas a la herramienta.

Por ello, como línea de estudio futuro, se propone un proyecto similar al actual, pero apoyándose en las herramientas de Jaspersoft, que como se ha podido observar en las comparativas, está muy a la par de Pentaho.

Por otro lado, aunque en un principio se hayan utilizado las estadísticas publicadas por el Ministerio de Educación como fuente de datos de apoyo, se han conseguido desarrollar informes de gran utilidad. Tal es su valor, que actualmente está en estudio su propuesta como propiedad intelectual de la Universidad Carlos III de Madrid.

11 REFERENCIAS

[01] *Quadrant for Business Intelligence and Analytics Platforms. 23 Febr. 2015* Disponible [Internet] http://www.gartner.com/technology/reprints.do?id=1-2ACLP1P&ct=150220&st=sb [10 mayo 2015]

[1] *Herramientas de Business Intelligence en el mercado*. Disponible [Internet]: http://sapuniverse.blogspot.com.es/2014/09/herramientas-de-business-intelligence.html [3 de Mayo de 2015]

[2] *Business Intelligence Software, BI with Analytics. SAP*. Disponible [Internet]: http://www.sap.com/spain/pc/analytics/business-intelligence.html [3 de Mayo de 2015]

[3] *IBM - Software Cognos – España.* Disponible [Internet]: http://www-01.ibm.com/software/es/analytics/cognos/ [4 de Mayo de 2015]

[4] *BI - Business Intelligence. Oracle.* Disponible [Internet]: http://www.oracle.com/us/solutions/business-analytics/business-intelligence/overview/index.html [4 de Mayo de 2015]

[5] *MicroStrategy: Analytics, Movilidad y Seguridad corporativas*. Disponible [Internet] http://www.microstrategy.com/es [4 de Mayo de 2015]

[6] *Business Intelligence in Office and SQL Server. Microsoft.* Disponible [Internet] http://www.microsoft.com/es-es/server-cloud/solutions/business-intelligence/ [4 de Mayo de 2015]

[7] *Pentaho - Data Integration, Business Analytics and Big Data*. Disponible [Internet] www.pentaho.com/ [4 de Mayo de 2015]

[8] *Jaspersoft Business Intelligence Software*. Disponible [Internet] https://www.jaspersoft.com/es [4 de Mayo de 2015]

[9] *Análisis e inteligencia de negocios. Tableau Software*. Disponible [Internet] http://www.tableau.com/es-es [4 de Mayo de 2015]

[10] *Qlik: Software de Business Intelligence y Visualización de Datos.* Disponible [Internet] http://global.qlik.com/es [4 de Mayo de 2015]

[11] *VMware Workstation.* Disponible [Internet] https://www.vmware.com/es/products/workstation [15 de Mayo de 2015]

[12] *Ubuntu 12.04.5 LTS (Precise Pangolin).* Disponible [Internet] http://releases.ubuntu.com/12.04/ [15 de Mayo de 2015]

[13] *The CPAN Search Site*. Disponible [Internet] search.cpan.org/ [15 de Mayo de 2015]

[14] *Data::PcAxis.* Disponible [Internet] http://search.cpan.org/~fod/Data-PcAxis-0.0.6/lib/Data/PcAxis.pm [15 de Mayo de 2015]

[15] *MYSQL*. Disponible [Internet] https://www.mysql.com/ [15 de Mayo de 2015]

[16] *Emma: Asistente para el manejo de MySQL*. Disponible [Internet] https://apps.ubuntu.com/cat/applications/emma/ [15 de Mayo de 2015]

[17] *Java EE 7 SDK*. Disponible [Internet] http://www.oracle.com/technetwork/java/javaee/downloads/java-ee-sdk-7-downloads-1956236.html [15 de Mayo de 2015]

[18] *Data Integration. Pentaho Community*. Disponible [Internet]. http://community.pentaho.com/projects/data-integration/

[19] *Reporting. Pentaho Community.* Disponible [Internet]. http://community.pentaho.com/projects/reporting/ [15 de Mayo de 2015]

[20] *Pentaho BI Server 5.0.* Disponible [Internet] http://community.pentaho.com/projects/data-integration/ [15 de Mayo de 2015]

[21] Manifiesto Ágil. Disponible [Internet] http://www.agilemanifesto.org/iso/es/principles.html [1 de Junio de 2015]

[22] *Custom Label Background in Pie Chart in Pentaho Report Designer*. Disponible [Internet]: https://bineedsui.wordpress.com/2015/05/25/custom-label-background-in-pie-chart-in-pentaho-report-designer/ [6 de julio de 2015]

[23] *X-Axis Label Wrap in Pentaho Report Designer*. Disponible [Internet]: https://bineedsui.wordpress.com/2013/12/23/x-axis-label-wrap-in-pentaho-report-designer/ [7 de julio de 2015]

[24] *Instituto Nacional de Estadística*. Disponible [Internet]: http://www.ine.es/ [1 de Marzo de 2015]

[25] *Estadísticas Gobierno Vasco*. Disponible [Internet]: www.eustat.eus/ [1 de Marzo de 2015]

[26] *Estadísticas Gobierno Canarias*. Disponible [Internet]: http://www.gobiernodecanarias.org/istac/ [1 de Marzo de 2015]

[27] *Estadísticas Gobierno de Aragón*. Disponible [Internet]: http://www.aragon.es/DepartamentosOrganismosPublicos/Institutos/InstitutoAragonesEstadistica/AreasGenericas/ci.Difusion_PcAxis.detalleTema [1 de Marzo de 2015]

[28] *Estadísticas Gobierno Balear*. Disponible [Internet]: http://ibestat.caib.es/ibestat/pcaxis [1 de Marzo de 2015]

[29] *Estadísticas Poder Judicial*. Disponible [Internet]: http://www.poderjudicial.es/cgpj/es/Servicios/Utilidades/Estadistica-Judicial-en-PC-AXIS [1 de marzo de 2015]

[30] *Estadísticas Gobierno Dinamarca*. Disponible [Internet]: http://www.dst.dk/en/OmDS/omweb/PC-AXIS# [1 de marzo de 2015]

[31] *Estadísticas Gobierno Suecia.* Disponible [Internet]: http://www.scb.se/sv_/PC-Axis/Start/ [1 de marzo de 2015]

[32] *Estadísticas Gobierno de China*. Disponible [Internet]: http://statdb.dgbas.gov.tw/pxweb/dialog/statfile1L.asp [1 de marzo de 2015]

[33] *INE. Estadísticas fabricación de coches*. Disponible [Internet]: http://www.ine.es/jaxi/menu.do?type=pcaxis&path=/t38/bme2/t05/b122&file=pcaxis [10 de septiembre de 2015]

[34] *INE. Precios Industriales.* Disponible [Internet] http://www.ine.es/dynt3/inebase/es/index.html?padre=632&dh=1 [10 de septiembre de 2015]

[35] *Ministerio de Educación. Estadísticas de estudiantes*. Disponible [Internet]: http://www.mecd.gob.es/educacion-mecd/areas-educacion/universidades/estadisticas-informes/estadisticas/alumnado.html [1 de marzo de 2015]

[36] *Pentaho Analysis Services (Mondrian).* Disponible [Internet]: http://sourceforge.net/projects/mondrian/files/ [7 de abril de 2015]

[37] *Pentaho Big Data*. Disponible [Internet]: http://www.pentaho.com/product/big-data-analytics [16 de junio de 2015]

[38] *Pentaho Data Mining. Weka.* Disponible [Internet]: http://community.pentaho.com/projects/data-mining/ [13 de julio de 2015]

[39] *Pentaho Metadata Editor*. Disponible [Internet]: http://wiki.pentaho.com/display/COM/The+Pentaho+Metadata+Project [15 de Mayo de 2015]

[40] *Pentaho Design Studio.* Disponible [Internet]: http://wiki.pentaho.com/display/ServerDoc2x/Design+Studio [15 de Mayo de 2015]

[41] *Ctools de Pentaho.* Disponible [Internet]: http://www.webdetails.pt/ctools/ [8 de septiembre 2015]

[42] *Saiku Reporting para Pentaho.* Disponible [Internet]: http://mgiepz.github.io/saiku-reporting/ [10 de septiembre de 2015]

12 BIBLIOGRAFÍA

- García Mattío, Mariano; R. Bernabeu, Dario. *Pentaho 5.0 Reporting By Example: Beginner's Guide.*

- *Community Wiki Home*. Disponible [Internet] http://wiki.pentaho.com [7 de Junio de 2015]

- *Dataprix.* Disponible[Internet] http://www.dataprix.com [15 de junio de 2015]

- El Rincón del BI. Disponible [Internet] https://churriwifi.wordpress.com [5 de Julio de 2015]

13 ANEXO

13.1 GUÍA DE OPERACIÓN

Cada cierre de curso académico el Ministerio de Educación sube los datos estadísticos del periodo en la siguiente web:

http://www.mecd.gob.es/servicios-al-ciudadano-mecd/estadisticas/educacion.html

Este proyecto se ha centrado en las estadísticas universitarias, concretamente en el número de alumnos matriculados.

Para actualizar el DWH con la última información subida es necesario seguir los siguientes pasos:

1. Entrar en el último periodo publicado.

Ilustración 47: Manual de Operación. Paso 1

2. Los ficheros necesarios están en el capítulo III, tanto para universidades públicas como privadas.

Ilustración 48: Manual de Operación. Paso 2

3. El objetivo es seleccionar la información más detallada posible, por lo que sólo se descargan los puntos b), d), e), y f), haciendo click en el icono "PAX" .

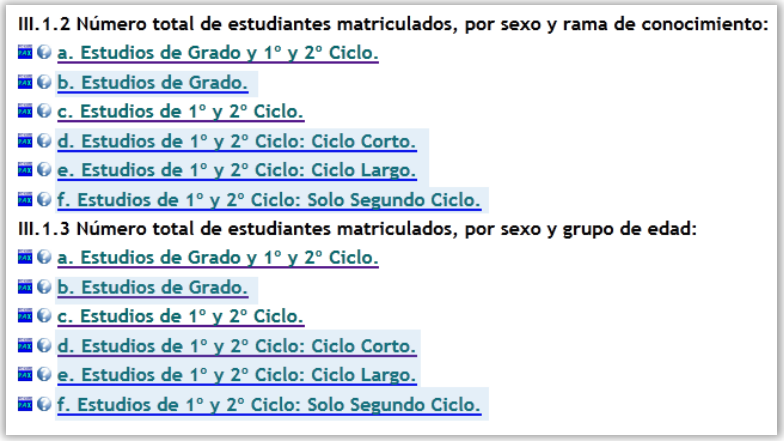


Ilustración 49: Manual de Operación. Paso 3

4. Los ficheros .px se descargan en el siguiente directorio del servidor Pentaho:

   /home/pentaho/pentaho-files/srcfiles

   Es recomendable renombrar los ficheros descargados, añadiendo el periodo al principio, para que sean más localizables (por ejemplo, "20152016_III3d_pri.px"). Además, es importante recordar el nombre, porque éste debe incluirse en un fichero de metadatos localizado en el mismo directorio llamado "ff_metadatos.txt"

5. Editar el fichero "ff_metadatos.txt".
   Es un fichero de texto, separado por tabulaciones que tiene los metadatos necesarios para cargar los datos estadísticos en el DWH. Su estructura es la siguiente (en el mismo orden):

   - Periodo
     Debe seguir el siguiente formato: "2009/2010"
   - Ciclo
     Los posibles valores pueden ser:

     "Matrícula de Grado"
     "1º y 2º ciclo. Ciclo Corto"
     "1º y 2º ciclo. Ciclo Largo"
     "1º y 2º ciclo. Solo segundo Ciclo"

   - Tipo Universidad

     "Pública" o "Privada"

- NULL (espacio reservado para futuras necesidades)
- Fichero entrada
  Nombre del fichero descargado, por ejemplo,

  "20152016_III3d_pri.px".

- Fichero de salida
  Si el fichero se ha descargado el punto 2, es decir, "Número total de estudiantes matriculados, por sexo y rama de conocimiento", el valor de este campo es:

  "ff_rama.txt"

  Si por el contrario, se ha descargado del punto 3, "Número total de estudiantes matriculados, por sexo y grupo de edad", el valor de este campo debe ser:

  "ff_edad.txt"

- Juego de Caracteres
  Aunque por regla general, casi todos los ficheros tienen el juego de caracteres cp437, algunas veces suben ficheros con otra codificación, por lo que los caracteres especiales podrían cargarse errónemanete en el DWH. Para evitar esto, una vez descargados todos los ficheros se ejecuta el comando:

  file –bi /home/pentaho/pentaho-files/srcfiles/<nombre del fichero>

  Si el comando devuelve el siguiente texto, la codificación es cp437:

  text/plain; charset=unknown-8bit

  Si devuelve este texto, la codificación es iso-8859-1:

  text/plain; charset=iso-8859-1

  Cualquier otro caso, debería ser estudiado ya que no se ha dado nunca hasta ahora.

6. El proceso que carga las estadísticas universitarias es incremental, por lo que podrían dejarse en el mismo directorio los ficheros de años anteriores.
7. Como último paso, se ejecuta el job j_estadisticas_academicas.

13.2 SCRIPTS PARA CUSTOMIZAR GRÁFICOS DE INFORMES

La herramienta Pentaho Report Designer tiene algunas carencias de diseño en los gráficos. Para suplir esta carencia, tiene la opción de incorporar scripts BeanShell, haciendo que su diseño sea totalmente configurable.

BeanShell es un lenguaje de scripting basado en la sintaxis de Java. Usa la máquina virtual Java y puede usar todas las librerías disponibles.

Estos scripts se definen dentro de la configuración de cada gráfico, en la sección “Scripting”:

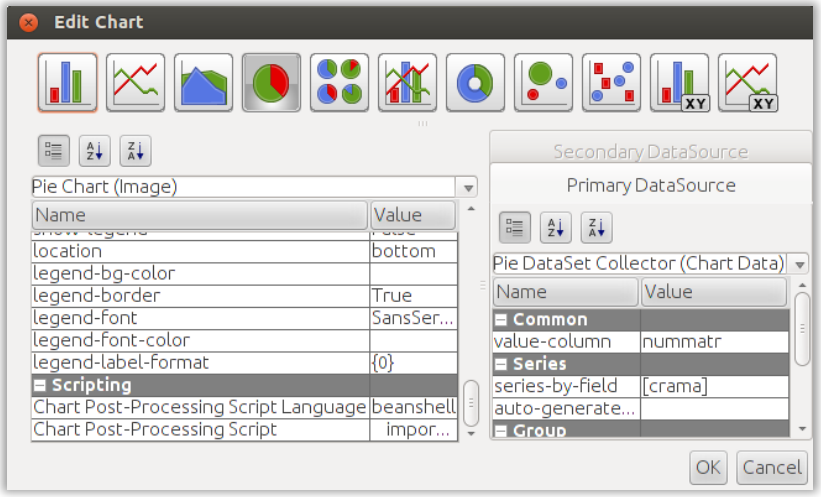


Ilustración 50: Scripts BeanShell para customizar gráficos

13.2.1 Modificar las etiquetas de los gráficos de tarta

Las etiquetas de los gráficos de tarta de Pentaho Report Designer tienen un aspecto poco estético, con fondo amarillo y borde. Para modificarlo se utiliza el siguiente BeanShell[22] :

```
import java.awt.Color;
import org.jfree.chart.plot.PiePlot3D;
import org.jfree.chart.plot.PiePlot;

PiePlot plot = (PiePlot) chart.getPlot();
plot.setLabelBackgroundPaint(Color.WHITE);
plot.setLabelOutlineStroke(null);
plot.setLabelShadowPaint(Color.WHITE);
```

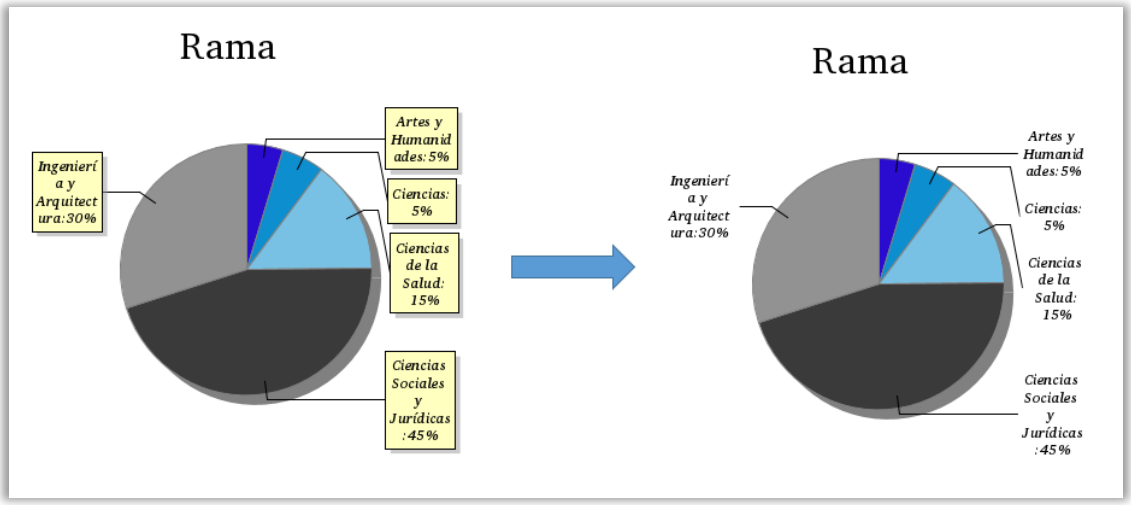


Ilustración 51: Resultado de BeanShell sobre gráficos de tarta

13.2.2 Ajustar etiquetas en gráfico de barras

Pentaho Report Designer trunca el texto de las etiquetas para adaptarlas al tamaño del gráfico, y no existe ninguna forma de configurar su tamaño, por lo que es muy posible que el resultado no sea el deseado.

Esto se puede solucionar con el siguiente script en BeanShell[23]:

```
import org.jfree.chart.plot.CategoryPlot;
import org.jfree.chart.axis.CategoryAxis;

CategoryPlot myPlot = chart.getCategoryPlot();
CategoryAxis myAxis = myPlot.getDomainAxis();
myAxis.setMaximumCategoryLabelLines(2);
```

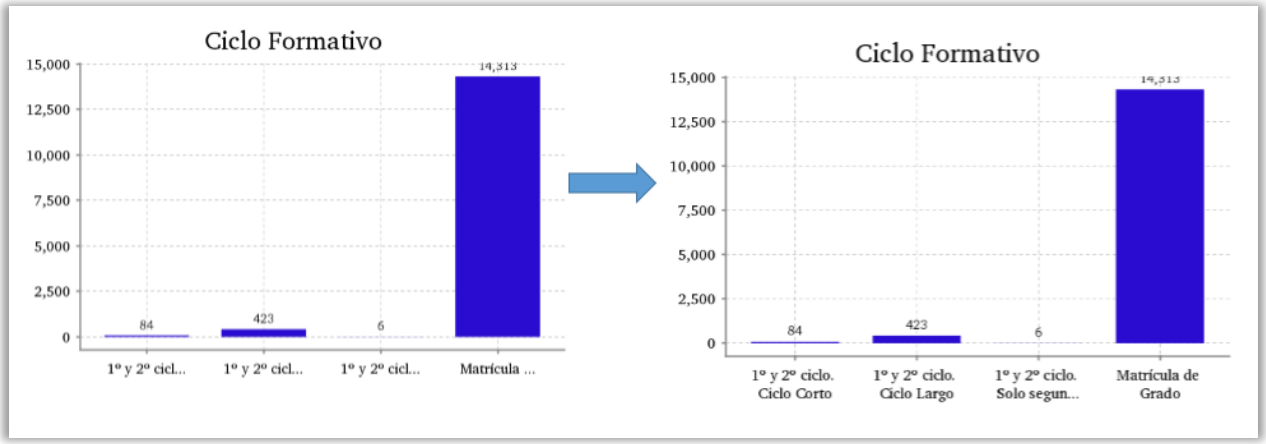


Ilustración 52: Resultado de BeanShell sobre gráficos de barras

13.3 SCRIPTS PERL

13.3.1 pcaxis2relational.pl

```
#!/usr/bin/perl
use Data::PcAxis;
use strict;

# Constructor

my ($fileinput, $fileoutput) = @ARGV;

my $px = Data::PcAxis->new($fileinput);

my $filename = $fileoutput;
open(my $fh, '>', $filename) or die "Could not open file '$filename' $!";

my $posdatum = 0;

sub pintar {
my ($jvariables, @pila) = @_;


my $num_vars  = $px->variables;

my $num_elemens = @pila;


if ($num_elemens < $num_vars )
{

for (my $j = $jvariables; $j < $num_vars; ++$j) {

my @val_names = $px->codes_by_idx($j);
my ($x,$y);

foreach $x (0..@val_names-1){
foreach $y (0..@{$val_names[$x]}-1){
my $num_elemens2 = @pila;
if ($num_elemens2 < $num_vars )
{
push @pila, $val_names[$x][$y];
pintar($j+1, @pila);
pop @pila;
}
}
}
}
}
else
{
   print $fh join("\t", @pila);
   my $res = $px->data->[$posdatum];
```

```
   print  $fh "\t$res\n";
   $posdatum ++;
}
}

print $fh "-----------------INICIO-------------------------\n";
my @val_names = $px->codes_by_idx(1);
my @elementos = ();
pintar(0, @elementos);

close $fh;
```

13.3.2 launcher.pl

```
#!/usr/bin/perl
use Data::PcAxis;
use strict;
use warnings;
use utf8;

# Constructor

my ($filemeta, $dirinput, $diroutput) = @ARGV;
open(my $fmeta, '<:encoding(utf8)', $filemeta) or die "Could not open file '$filemeta' $!\n";

#while (my $linemeta = 0){
while (my $linemeta = <$fmeta>){
chomp $linemeta;
$linemeta =~ s/ /+++/g;

my @fields = split "\t", $linemeta;

my $encode = pop @fields;
my $fileoutput = join '', $diroutput, pop @fields;
my $fileinput =  join '', $dirinput, pop @fields;
my $filencodig = join '', $fileinput, 'new';

my $textencoding =  "<:encoding($encode)";

open my $filascii, "<:encoding($encode)", $fileinput;
open my $fileutf8, '+>:encoding(utf8)', $filencodig;
print $fileutf8 $_ while <$filascii>;

close $filascii;
close $fileutf8;

rename ($filencodig, $fileinput) || die "Cannot rename: $!";

qx(perl /home/pentaho/pentaho-files/scripts/perl/pcaxis2relational_academic.pl $fileinput  $fileoutput @fields);
}
```

13.4 DDL O DEFINICIÓN DE DATOS

A continuación se describen los scripts de generación de las estructuras que conformen la BBDD. El gestor de Bases de Datos utilizado es MySQL.

```
CREATE TABLE ‘deaciclo’ (
  ‘idciclo’ int(11) NOT NULL,
  ‘cciclo’ varchar(250) NOT NULL,
  ‘dfeccre’ date NOT NULL,
  PRIMARY KEY (‘idciclo’)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COMMENT='latin1_swedish_ci'


CREATE TABLE ‘deaedad’ (
  ‘idrangoedad’ int(11) NOT NULL,
  ‘crangoedad’ varchar(250) NOT NULL,
  ‘dfeccre’ date DEFAULT NULL,
  PRIMARY KEY (‘idrangoedad’)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COMMENT='latin1_swedish_ci'


CREATE TABLE ‘deagenero’ (
  ‘idgenero’ int(11) NOT NULL,
  ‘cgenero’ varchar(250) NOT NULL,
  ‘dfeccre’ date NOT NULL,
  PRIMARY KEY (‘idgenero’),
  KEY ‘idx_destacagenero_lookup’ (‘cgenero’)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COMMENT='latin1_swedish_ci'


CREATE TABLE ‘deanacionalidad’ (
  ‘idnacionalidad’ int(11) NOT NULL,
  ‘cnacionalidad’ varchar(250) NOT NULL,
  ‘dfeccre’ date DEFAULT NULL,
  PRIMARY KEY (‘idnacionalidad’)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COMMENT='latin1_swedish_ci'

CREATE TABLE ‘dearama’ (
  ‘idrama’ int(11) NOT NULL,
  ‘crama’ varchar(250) NOT NULL,
  ‘dfeccre’ date DEFAULT NULL,
  PRIMARY KEY (‘idrama’)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COMMENT='latin1_swedish_ci'


CREATE TABLE ‘deatipuniv’ (
  ‘idtipuniv’ int(11) NOT NULL,
  ‘ctipuniv’ varchar(250) NOT NULL,
  ‘dfeccre’ date NOT NULL,
  PRIMARY KEY (‘idtipuniv’)
```

```
) ENGINE=InnoDB DEFAULT CHARSET=latin1


CREATE TABLE ‘deauniv’ (
  ‘iduniv’ int(11) NOT NULL,
  ‘cuniv’ varchar(250) NOT NULL,
  ‘idtipuniv’ int(11) NOT NULL DEFAULT '-1',
  ‘dfeccre’ date DEFAULT NULL,
  PRIMARY KEY (‘iduniv’)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COMMENT='latin1_swedish_ci'


CREATE TABLE ‘aeaedad’ (
  ‘cperiodo’ char(9) NOT NULL,
  ‘iduniv’ int(11) NOT NULL,
  ‘idgenero’ int(11) NOT NULL,
  ‘idciclo’ int(11) NOT NULL,
  ‘idrangoedad’ int(11) NOT NULL,
  ‘nnummatriculados’ decimal(13,2) NOT NULL,
  PRIMARY KEY (‘cperiodo’,’iduniv’,’idgenero’,’idciclo’,’idrangoedad’)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COMMENT='latin1_swedish_ci'

CREATE TABLE ‘aeanacionalidad’ (
  ‘cperiodo’ char(9) NOT NULL,
  ‘iduniv’ int(11) NOT NULL,
  ‘idgenero’ int(11) NOT NULL,
  ‘idciclo’ int(11) NOT NULL,
  ‘idnacionalidad’ int(11) NOT NULL,
  ‘nnummatriculados’ decimal(13,2) NOT NULL,
  PRIMARY KEY (‘cperiodo’,’iduniv’,’idgenero’,’idciclo’,’idnacionalidad’)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COMMENT='latin1_swedish_ci'


CREATE TABLE ‘aearama’ (
  ‘cperiodo’ char(9) NOT NULL,
  ‘iduniv’ int(11) NOT NULL,
  ‘idgenero’ int(11) NOT NULL,
  ‘idciclo’ int(11) NOT NULL,
  ‘idrama’ int(11) NOT NULL,
  ‘nnummatriculados’ decimal(13,2) NOT NULL,
  PRIMARY KEY (‘cperiodo’,’iduniv’,’idgenero’,’idciclo’,’idrama’)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COMMENT='latin1_swedish_ci'
```